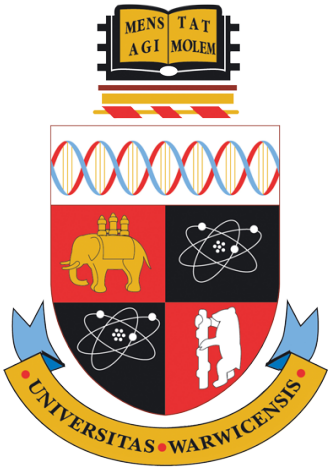


Machine learning with limited information: risk stratification and predictive modelling for clinical applications

by

Torgyn Shaikhina

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

School of Engineering

September 2017

Declaration of authorship

I, Torgyn Shaikhina, declare that the work presented in this thesis is my own and was carried out in the School of Engineering, the University of Warwick during the period from October 2013 to September 2017. Parts of this thesis have been published in several peer-reviewed journals during the course of this Ph.D. programme, as detailed in the list of Publications and referenced in the text. The research presented here has not been submitted in whole or in part for any degree at this or any other university.

Acknowledgements

Foremost, I would like express profound gratitude to my supervisor, Dr. Natasha Khovanova, who kindled my passion for academic research, tirelessly encouraged and guided me, taught me the scientific method, co-authored our papers, involved me in her clinical collaborations, and gave me numerous opportunities for dissemination of the results and personal development. I would also like to thank my late supervisor, Dr. Kajal Mallick, and my second supervisor, Dr. Mark Leeson, for his ideas and wisdom.

It is an honour to acknowledge the nephrology collaborators: Prof. Robert Higgins and Prof. Nithya Krishnan (University Hospitals Coventry and Warwickshire), Prof. David Briggs and Dr. David Lowe (NHS Blood and Transplant Birmingham), Dr. Sunil Daga (Leeds Teaching Hospitals), and Dr. Daniel Mitchell (Warwick Medical School). Working with them added practical depth and a meaning to my research.

I am grateful to the second team of clinical and statistical collaborators at the Nuffield Department of Primary Care Health Sciences (University of Oxford): Dr. Tim Holt, Margaret Smith, Alice Fuller, Dr. Claire Bankhead, Sarah Stevens, and Prof. Rafael Perera, for their enthusiasm in machine learning and contagious passion for primary care.

I would like to thank my mentor, Dr. James Covington, for his continued support throughout my years at Warwick, and all my colleagues and friends at the School of Engineering. A special mention goes to Katherine Heathward for her commitment to finding every misplaced article in this thesis and beyond.

Finally, this work would have been unsurmountable without the support and encouragement that I received from my beloved husband David and the rest of my family and friends during the course of this Ph.D. programme.

Publications

Results of this research have been published in 8 peer-reviewed papers and disseminated at 15 conferences and invited talks, as detailed below:

List of papers

1. Khovanova N., Shaikhina T., Mallick K. (2014) "Artificial Neural Network approach to multidimensional correlation analysis of a trabecular bone", *Bioinspired, Biomimetic and Nanobiomaterials.* 4 (1), 90-100

2. Shaikhina T., Khovanova N., Mallick K. (2014) "Artificial neural networks in hard tissue engineering: another look at age-dependence of trabecular bone properties in osteoarthritis", *Proceedings of IEEE-EMBS International Conference on Biomedical and Health Informatics*, 622-625.

3. Khovanova N., Daga S., Shaikhina T., Krishnan N., Jones J., Zehnder D., Mitchell D., Higgins R., Briggs D. and Lowe D. (2015) "Subclass analysis of donor HLA-specific IgG in antibody-incompatible renal transplantation reveals a significant association of IgG4 with rejection and graft failure", *Transplant International*, 28 (12), 1405-1415

4. Shaikhina T., Khovanova N., Daga S., Krishnan N., Lowe D., Mitchell D., Briggs D., Higgins R. (2015) "Prediction of acute antibody mediated rejection in antibody incompatible renal transplantation using machine learning for wide data", *American Journal of Transplantation*, 15 (S3), 1364 – *short paper*

5. Khovanova N., Lowe D., Daga S., Shaikhina T., Mitchell D., Zehnder D., Briggs D., Higgins R. (2015) "Assessment of IgG subclass significance for early graft rejection and long-term survival in HLA-antibody incompatible renal transplantation: multivariate approach", *American Journal of Transplantation*, 15 (S3), 338 – *short paper*

6. Shaikhina T., Lowe D., Daga S., Briggs D., Higgins R., Khovanova N. (2015) "Machine learning for predictive modelling based on small data in biomedical engineering", *IFAC-PapersOnLine*, 48(20), 469-474

7. Shaikhina T., Khovanova N. (2017) "Handling limited datasets with neural networks in medical applications: A small-data approach", *Artificial Intelligence in Medicine*, 75, 51-63

8. Shaikhina T., Lowe D., Daga S., Briggs D., Higgins R., Khovanova N. (2017) "Decision trees and random forests models for outcome prediction in high risk kidney transplantation", *Biomedical Signal Processing and Control – In press (DOI: doi.org/10.1016/j.bspc.2017.01.012)*

9. Shaikhina T., Smith M., Khovanova N., Leeson M., Fuller A., Bankhead C., Stevens S., Perera R., Holt T. (2017) "Identifying risk of diabetes in primary care using an artificial neural network: comparative cohort study" – *In preparation*

Conference disseminations

1. Shaikhina T., Khovanova N., Mallick K. "Neural network modelling of trabecular bone for hard tissue engineering", *8th Workshop for Women in Machine Learning (WiML)*, Lake Tahoe, Nevada, USA, 5 Dec 2013 – *poster, NIPS travel grant*

2. Shaikhina T., Khovanova N., Mallick K. "Artificial neural networks in hard tissue engineering: another look at age-dependence of trabecular bone properties in osteoarthritis", *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Valencia, 1-4 Jun 2014 – *oral presentation*

3. Shaikhina T., Khovanova N. "Predicting mechanical strength of trabecular bone in severe osteoarthritis using neural network", *[id]2$_{Ox}$ inter-disciplinary inter-DTC Student Conference*, Oxford, 26-27 Jun, 2014 – *oral presentation* (**Best Oral Presentation Prize by AstraZeneca**)

4. Shaikhina T., Khovanova N. "Application of machine learning to bioscaffold design decision support in hard tissue engineering", *8th IEEE EMBS International UK & Republic of Ireland Postgraduate Conference on Biomedical Engineering and Medical Physics*, Warwick University, 15-17 Jul 2014 – *oral presentation, session chair, member of organising committee*

5. Shaikhina T., Daga S., Krishnan N., Lowe D., Mitchell D., Briggs D., Higgins R., Khovanova N. "Decision trees in renal transplantation: prediction of acute antibody mediated rejection in the early post-transplant period", *3rd International Transplantation conference: How much risk can you take (and what to do if it all goes pear shaped)*, Coventry, 31 Oct -1 Nov 2014 – *oral presentation*

   Khovanova N., Lowe D., Daga S., Shaikhina T., Krishnan N., Zehnder D., Briggs D., Higgins R. "Presence and levels of non-complement fixing IgG4 subclass associates with early graft rejection and decreased allograft survival times in antibody incompatible transplantation", *3rd International Transplantation conference: How much risk can you take and what to do if it all goes pear shaped*, Coventry, 31 Oct - 1 Nov 2014 – *contribution*

6. Shaikhina T., Khovanova N. "Neural networks as a prediction tool for small experimental datasets in biomedical engineering", *9th Annual Workshop for Women in Machine Learning (WiML)*, Montreal, 8 Dec 2014 – *poster, NIPS travel grant*

7. Shaikhina T., Khovanova N., Daga S., Krishnan N., Lowe D., Mitchell D., Briggs D., Higgins R. "Decision trees for small data sets: prediction of acute antibody mediated rejection in early post-transplant period in antibody incompatible transplantation", *Joint British Transplantation Society and Nederlandse Transplantatie Vereniging Congress*, Bournemouth, 11-13 Mar 2015 – *poster*

   Khovanova N., Lowe D., Daga S., Shaikhina T., Krishnan N., Mitchell D., Zehnder D., Briggs D., Higgins R. "IgG4 subclass associates with early graft rejection and decreased allograft survival times in antibody incompatible transplantation", *Joint British Transplantation Society and Nederlandse Transplantatie Vereniging Congress*, Bournemouth, 11-13 Mar 2015 – *contribution*

8. Shaikhina T., Khovanova N., Daga S., Krishnan N., Lowe D., Mitchell D., Briggs D., Higgins R. "Prediction of acute antibody mediated rejection in antibody incompatible renal transplantation using machine learning for wide data", *American Transplant Congress*, Philadelphia, USA, 2 - 6 May 2015 – *oral presentation*

   Khovanova N., Lowe D., Daga S., Shaikhina T., Mitchell D., Zehnder D., Briggs D., Higgins R. "Assessment of IgG subclass significance for early graft rejection and long-term survival in HLA-antibody incompatible renal transplantation: multivariate approach", *American Transplant Congress*, Philadelphia, USA, 2 - 6 May 2015 – *contribution*

9. Shaikhina T., Lowe D., Daga S., et al. "Machine learning for predictive modelling based on small data in biomedical engineering", *9th IFAC Symposium on Biological and Medical Systems*, Berlin, 2-5 Sep 2015 – *oral presentation*

10. Shaikhina T., Khovanova N. "Making sense of complex clinical data using machine learning", *SET for Britain*, *Engineering and Mathematical Sciences Exhibition*, Westminster, London, 7 Mar 2016 – *poster* (**shortlisted by the Parliamentary and Scientific Committee in collaboration with the Royal Academy of Engineering, and invited to present in the House of Commons**)

11. Babu A., Andreou A., Porumb M., Shaikhina T., Briggs D., Krishnan N., Barber T., Mitchell D., Higgins R., Khovanova N., Daga S. "Presence of day-14 post transplantation donor specific IgM antibody predicts poor graft survival in HLA-incompatible renal transplantation" and "Clinical relevance of pre-formed IgM HLA-donor specific antibodies (DSA) in HLA-incompatible kidney transplantation", *26th International Congress of the Transplantation Society (TTS)*, Hong Kong, 18-25 Aug 2016 – *contributions*

12. Shaikhina T., Daga S., Krishnan N., Lowe D., Mitchell D., Briggs D., Higgins R., Khovanova N. "Ensemble Learning for clinical data analysis and outcome prediction in Renal Transplantation: Random Forests", *4th International Transplant Conference*, Coventry, 25-26 Nov 2016 – *oral presentation*

13. Shaikhina T., Smith M., Khovanova N., Leeson M., Fuller A., Bankhead C., Stevens S., Perera R., Holt T. "Identifying risk of diabetes in primary care using an artificial neural network: comparative cohort study", *National Institute for Health Research SPCR Showcase 2017*, Blavatnik School of Government, Oxford, 19 Sept 2017 – *oral presentation*

Invited talks

1. "Machine Learning for clinical data analysis, risk stratification and outcome prediction in Renal Transplantation", *King's College London MRC Centre for Transplantation Lunchtime Seminar*, London, 8 Apr 2016

2. "Machine Learning for predictive modelling based on small biomedical and clinical data: From concrete to bones", *Warwick Centre for Predictive Modelling Seminar Series*, Coventry, 10 May 2016

Grants and scholarships

1. *Chancellor's International Scholarship* (£96,000, Mar 2013) for 3.5 years of Ph.D. funding including overseas tuition fees and maintenance stipend, University of Warwick, UK

2. Two *Women in Machine Learning student presenter* travel scholarships to attend Neural Information Processing Systems conference in the USA ($800, Oct 2013) and Canada ($1000, Oct 2014)

3. *Monash Warwick Alliance Student-led Activity Fund grant* (£12,400, Dec 2014) for co-founding and organizing Machine Learning Bootcamp, Melbourne, Australia.

4. *Next Generation Women Leaders Award* (2,000 EUR, Jan 2016) recognising individuals with strong leadership potential, teamwork and problem-solving by McKinsey & Co., London, UK

5. *Lord Rootes Memorial Award* (£5,500, Apr 2016) for founding Next Generation Programmers initiative for rural developing countries, Bayanaul, Kazakhstan

Abstract

The high cost, complexity and multimodality of clinical data collection restrain the datasets available for predictive modelling using machine learning (ML), thus necessitating new data-efficient approaches specifically for limited datasets. This interdisciplinary thesis focuses on clinical outcome modelling using a range of ML techniques, including artificial neural networks (NNs) and their ensembles, decision trees (DTs) and random forests (RFs), as well as classical logistic regression (LR) and Cox proportional hazards (Cox PH) models. The utility of ML for data-efficient regression, classification and survival analyses was investigated in three clinical applications, whereby exposing the common limitations inherent in patient data, such as class imbalance, incomplete samples, and, in particular, limited dataset size. The latter problem was addressed by developing a methodological framework for learning from datasets with less than 10 observations per predictor variable. A novel method of multiple runs overcame the volatility of NN and DT models due to limited training samples, while a surrogate data test allowed for regression model evaluation in the presence of noise due to limited dataset size. When applied to hard tissue engineering for predicting femoral fracture risk, the framework resulted in 98.3% accurate regression NN. The framework was used to detect early rejection in antibody-incompatible kidney transplantation, achieving 85% accurate classification DT. The third clinical task – that of predicting 10-year incidence of type 2 diabetes in the UK population – resulted in 70-85% accurate classification and survival models, whilst highlighting the challenges of learning with the limited information characteristic of routinely collected data. By discovering unintuitive patterns, supporting existing hypotheses and generating novel insight, the ML models developed in this research contributed meaningfully to clinical research and paved the way for data-efficient applications of ML in engineering and clinical practice.

Dedicated to my Ata

Shaikhin Balga (1936–2006)

Contents

List of tables

List of figures

Abbreviations

ABMR	**A**nti**b**ody-**M**ediated **R**ejection
ADASYN	**Ada**ptive **Syn**thetic sampling
AIT	**A**ntibody-**I**ncompatible **T**ransplantation
AUC	**A**rea **u**nder the ROC **C**urve (see ROC)
BG	**B**lood **G**lucose
BMD	**B**one **M**ineral **D**ensity
BMI	**B**ody **M**ass **I**ndex
BV/TV	**B**one **V**olume over **T**otal **V**olume
CART	**C**lassification **a**nd **R**egression **T**rees
CDC	**C**omplement-**D**ependent **C**ytotoxic crossmatching
Cox PH	**Cox P**roportional **H**azards
CPRD	**C**linical **P**ractice **R**esearch **D**atalink
CS	**C**ompressive **S**trength
CT	**C**omputer **T**omography
CVD	**C**ardio**v**ascular **D**isease
DGF	**D**elayed **G**raft **F**unction
DM	**D**iabetes **M**ellitus
DSA	**D**onor-**S**pecific **A**ntibody
DT	**D**ecision **T**ree
EMR	**E**lectronic **M**edical **R**ecords
ESRD	**E**nd-**S**tage **R**enal **D**isease
FBG	**F**asting **B**lood **G**lucose
FEA	**F**inite-**E**lement **A**nalysis
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
GDI	**G**ini's **D**iversity **I**ndex

HLA	**H**uman **L**eukocyte **A**ntigen
HR	**H**azard **R**atio
IgG	**I**mmunoglobulin **G**
LR	**L**ogistic **R**egression
MAR	**M**issing **a**t **R**andom
MCAR	**M**issing **C**ompletely **at R**andom
MFI	**M**edian **F**luorescence **I**ntensity
MICE	**M**ultiple **I**mputation with **C**hained **E**quations
ML	**M**achine **L**earning
MNAR	**M**issing **n**ot **a**t **R**andom
MSE	**M**ean **S**quared **E**rror
NN	**N**eural **N**etwork
NHS	**N**ational **H**ealth **S**ervice (UK)
NPV	**N**egative **P**redictive **V**alue
PPV	**P**ositive **P**redictive **V**alue
RF	**R**andom **F**orest
ROC	**R**eceiver **O**perating **C**haracteristic curve
SCG	**S**caled **C**onjugate **G**radient
SMI	**S**tructure **M**odel **I**ndex
SMOTE	**S**ynthetic **M**inority **O**versampling **Te**chnique
Tb.Th	**T**ra**b**ecular **Th**ickness
TF	**T**ransfer **F**unction
TN	**T**rue **N**egative
TP	**T**rue **P**ositive
UHCW	**U**niversity **H**ospitals **C**oventry and **W**arwickshire

Symbols

Greek

β	Regression model parameters
e^{β}	Odds ratio / Hazard ratio
η	Number of neurons in the hidden layer
θ	Set of network parameters
λ	Hazard
λ_o	Baseline hazard
μ	Mean
σ	Standard deviation
χ^2	Chi-squared test value
ω	Early stopping criterion

Latin

b	Set of network biases
E	Error between predicted y and target t outputs
n	Number of observations (samples)
p	Number of predictor variables (features)
Q	Set of paired inputs X and targets t
r	Relative event rate of a survival tree
t	Vector of expected model targets
W_I	Matrix of input weights
w_L	Vector of layer weights
X	Input matrix with p features and n observations
x_j	One of p input variable vectors with n observations
y	Vector of predicted model outputs
$\emptyset$	Empty set

Performance measures

AUC	Area under the receiver operating characteristic curve
C	Correct classification rate
$C_{balanced}$	Balanced accuracy of classification
CE	Cross entropy
C-index	Harrell's concordance index
D	Royston and Sauerbrei's D score
LL	Log-likelihood
LP	Log partial likelihood
MSE	Mean squared error
R	Coefficient of determination (regression factor)
R_D^2	R^2 factor based on Royston and Sauerbrei's D
$RMSE$	Root mean squared error
Sn	Sensitivity
Sp	Specificity

Subsets of data

all	All samples
$model$	Training and validation samples
$surr$	Surrogate samples
$train$	Training samples
$test$	Testing samples
val	Validation samples

Chapter 1

Introduction

1.1 The tacit value of expert insight

In his 1946 study on human expertise, Dutch chess master and psychologist Adriaan de Groot came to a striking revelation that for a given position, grandmasters evaluated fewer moves than less experienced players, but each of those moves were among the five best possible. The grandmasters, he noted, were able to 'immediately "see" the core of the problem in the position' [1]. This ability of a practiced mind to eliminate poor solutions, before they reached the conscious thought, is attributed to *insight*. Defining *information* as value-added raw *data* in a usable context, and *knowledge* – as an interconnected and structured system of information (Figure 1.1), *insight* could be described as a sudden change in knowledge *representation,* often leading to a formulation of a new concept or awareness of a solution [2].

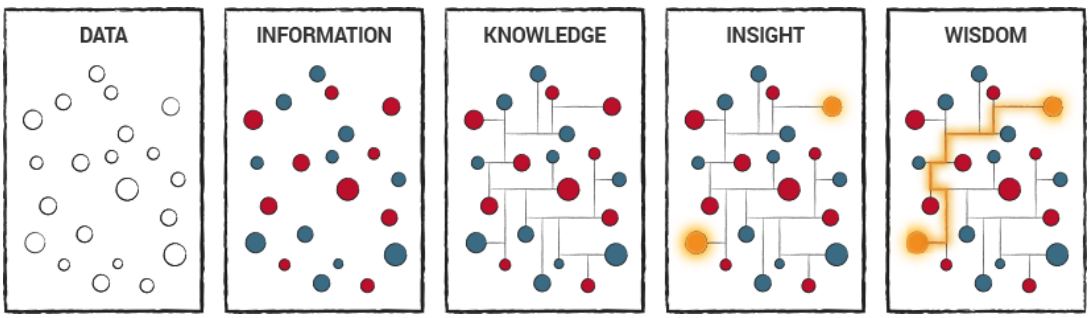


Figure 1.1 Graphical illustration of the relationship between data, information, knowledge, insight and wisdom. Adapted from [3].

For a society that has spent most of the 20th century meticulously codifying existing human knowledge into computer algorithms, insight presented a challenge: it is difficult to reproduce, systematise or even explain the tacit mechanisms by which we generate insight. Without statistical backing and reproducibility, insight is often viewed as intuition, creative genius, or luck, therefore, invalidating its use in systematic decision-making in high-risk applications and expert domains such as medicine [4]. Although largely a product of *subconscious* processing, insight originates from *learning complex patterns* in the data, that results in an often spontaneous awareness of the underlying knowledge structure [2]. Whether in chess or radiography, the more practice we undertake recognising relevant patterns, the more efficient and focused our thinking becomes, to the point where we are able to ignore non-essential knowledge pathways and come up with an insight (Figure 1.1).

Pattern recognition is a fundamental function of a human brain through which we integrate sensory information about the surrounding world to generate advantageous behavioural responses. Whether it is in detecting familiar faces and creating cognitive maps of the environment, or piecing together sounds into a language, throughout thousands of years of existence, humans have developed *a remarkable capacity for pattern recognition* from sensory inputs. We are even capable of identifying *temporal patterns* when, driven by our rudimentary aversion of uncertainty, we access past and present information to *predict and model* the future [5]. Nevertheless, our perception of *numerical* and *abstract patterns* is often limited by three-dimensional spatial thinking, making spontaneous insight improbable when multi-dimensional properties of observations are involved, such as that in identifying a rare disease from sporadic symptoms, unrelated tests and fragmented medical history.

1.2 Teaching machines to generate insight

From drawings and clay models to mathematics and physical theories – we have always used instruments to augment our perception of sensory, numerical, and abstract patterns and aide our *predictions.* Yet only with the advent of computers we have been able to overcome the limitation of our three-dimensional spatial thinking. This became possible as a result of *machine learning* (ML) – a paradigm in which computers are taught to learn patterns from data, as opposed to being pre-programmed with equations that describe these patterns. By teaching computers to recognise patterns in the data, we have been able to augment our pattern recognition in higher dimensions, provide statistical backing to our intuition, and make probabilistically-founded predictions from existing observations [6].

Formally, ML refers to an area of artificial intelligence that enables autonomous learning from input stimuli [7–9]. The process of learning in itself is a stepwise refinement of patterns that results from repeated hierarchical, parallel and recursive computations on new observations and experiences [4]. As with human learning, where we adapt our behaviour based on observed information in order to achieve a set goal, computers are trained to generate adaptive responses to the input data when given an objective function. This fundamental property allows ML systems to generate insight from complex patterns through exposure to data.

The ML domain encompasses a diverse array of algorithms and modes of learning, including supervised, unsupervised, hybrid, ensemble, reinforcement, active, adversarial, and transfer learning [8–12]. In *predictive modelling* that is the focal point of this thesis, much of the success of ML has been attributed to *supervised learning* [6,13], in which machines learn to map inputs to pre-specified desired outputs (ground truth).

By comparing ML system predictions with the true outcomes, supervised learning essentially reproduces the "trial and error" approach of human learning [4]. Once trained, the models can be used for *generalising* on new input data, i.e. the prediction of outputs for sets of data not previously encountered by the model. Supervised learning has proven exceptionally effective in solving problems that involve sorting sets of data into previously known classifications, mapping trends (function fitting), and forecasting the output from sets of inputs [14,15].

Among numerous ML architectures, *artificial neural networks* (NN) are widely recognised for their ground-breaking ability to derive insight from complex non-linear patterns. Originally inspired by the function of biological neurons in the central nervous system [16,17], NNs are regarded as universal function approximators capable of working with both linear and non-linear systems [18,19]. Various neural learning algorithms and network configurations have been developed throughout NN history [20], allowing NNs to be tailored to the demands of specific analytical tasks ranging from classification, forecasting and time series analysis to combinatorial problem solving, adaptive control, multisensory data fusion and noise filtering [21,22]. Most recently, Deep Learning [23] resulted in NNs mastering speech recognition [24] and surpassing humans in the game of Go [25]. Despite their adaptability, NN algorithms are sensitive to the quality and size of the training data. Precisely for this versatility, promising future potential, and realistic limitations, the NN was chosen in this research as the core ML architecture for exploring the challenges of predictive modelling with limited information.

In some high-risk clinical applications such as organ transplantation, where an erroneous decision can be fatal, the black-box nature of NNs render them inadmissible to critical decision support. In applications that require an intuitive model to be used by

the operating physician, *decision tree* (DT) learning offers an unprecedented transparency of the statistical pattern associations with transplantation outcomes.

Although largely focused on NNs and DTs, the methodological approaches developed in this thesis could be applied to other ML systems for clinical risk stratification, including kernel-based learning [318] and Bayesian inference models [200]. However, these systems are outside of the remit of this research and have not been considered in this thesis.

1.3 Machine learning in healthcare

In medicine and clinical epidemiology, ML is beginning to enable predictive decision support for healthcare professionals in diagnosing diseases [26], predicting mortality or relapse [27], informing treatment strategies [28–30], simulating potential outcomes [31,32], and in performing numerous other patient-specific analyses [33]. ML is viewed as an indispensable tool for biomedical problems involving complex heterogeneous data when conventional statistical tools fail [34–36]. In applications such as gene selection and classification [37], screening heart murmurs in children [38], and predicting breast cancer relapse [39], ML models were able to map highly nonlinear input and output patterns even when mechanistic relationships between model variables could not be determined due to pathologies or complexity.

ML is largely responsible for the recent breakthrough in human genomics [40,41] and drug discovery [42,43], thus accelerating our transition to *precision medicine* [44,45]. In radiology and brain imaging, computer-aided image recognition is reshaping the clinical practice and reaching above-human diagnostic accuracies [46,47]. In other areas, clinical researchers equipped with ML are working to eradicate AIDS [28,32], treat diverse types of cancer [29,48–50], and improve the efficiency of critical care [51]. Despite these

notable advances in medical *research*, the uptake of ML in clinical *practice* has been slower than in many other equally high-risk expert domains [52,53]. The vast potential of ML for predictive modelling in healthcare remains largely unexplored. It is argued that the further development of ML in clinical practice is impeded by limited availability and quality of relevant data [33,51,52]. Thus, to extend the benefits of ML to a wider range of clinical datasets, it is essential to first develop methods that would compensate for the *limited data*. This brings us to a consideration of the inherent properties of clinical and biomedical data that make ML particularly challenging.

1.4 Why clinical data are limited

The limitations of clinical datasets are two-fold: *limited availability of high quality data* and *low quality of available data.* The *size* of datasets available for statistical modelling is often restrained by the cost and complexity of medical experiments. Most experimental datasets stem from single-centre studies, and do not meet the demands of data-intensive ML systems designed for Big data [54]. Multicentre data collaborations have proven to possess a tremendous potential for transforming clinical practice [53], yet integration of datasets across multiple institutions remains problematic. Heterogeneity of study protocols, differing international and inter-institutional standards, concerns over patient confidentiality, and technical incompatibility – all pose barriers to an open sharing of medical information and the creation of the substantive datasets required to train ML systems. Finally, in some medical domains, for example, *hard tissue engineering* or *organ transplantation*, where associated interventions are invasive and potentially harmful to the patient, obtaining large number of samples is altogether unrealisable.

Routinely collected patient data, such as those found in *electronic medical records* (EMR) and clinical data management systems, are generally more accessible than data curated for research purposes. By reusing the standard clinical databases, it is possible to decrease the cost of large volume datasets for training ML systems. Nevertheless, the *quality* of routinely collected data is often limited in terms of:

- low information density
- multimodality and heterogeneity
- missing values and censoring
- class imbalance
- corruption by noise and errors

Firstly, not all data are informative (Figure 1.1). Even a large EMR database may not contain the necessary *information* to reliably infer complex patterns relevant to a clinical problem at hand. For instance, in order to successfully model a multifactorial disease such as diabetes mellitus, the data must carry sufficient information to describe the pathophysiological mechanisms of the disease, to differentiate between patient phenotypes, and to account for confounding factors and the reverse causality of diabetes with conditions such as hypertension and obesity. As later discovered in this research, such intricate detail is not presently available in the UK healthcare EMR systems. Some leading groups in clinical ML adaptation argue that commercial clinical systems were designed "to document clinical activity for reporting, liability, and billing reasons, rather than for developing new algorithms" [51]. The low information density of such EMR systems make even large datasets small.

Multimodality of a clinical dataset refers to its heterogeneous sources: free-text reports by general practitioners, hospital discharge records, biochemistry tests and biopsies from laboratories, DNA sequencing data, vaccination history, X-ray and MRI scans,

medication prescriptions, and even medical insurance claims. Multimodality poses similar problems with dataset integration as the multicentre data discussed earlier, but requires technologically different ML solutions that cascade or combine multiple learners [55–57]. Variable types in clinical datasets also exhibit heterogeneity: *discrete* (ex.: family size, number of previous transplants) and *continuous* (ex.: age, blood pressure, blood glucose level) *numerical* variables, *nominal* (ex.: gender, blood group) and *ordinal* (ex.: degrees of pain, classes of antibodies) *categorical* factors, including binary indicators, as well as image matrices, waveforms and time series – all of which necessitate a mixed modelling approach and non-trivial pre-processing [17,58] .

Clinical database records and EMR are *rarely complete*. Missing values arise in routinely collected datasets for reasons such as unsystematic recording, equipment failure, time constraints, human error, system blackouts, and patient no-show. Some unrecorded values may be implicit (e.g.: sex of obstetrics patients, ethnicity in homogeneous communities) or sparse by design (e.g.: historic data for a newly-introduced variable). A particular type of missing data are right-*censored* observations, where the outcome variable is unknown due to loss of the patient follow-up. Censored population data are particularly difficult for supervised learning; limited success has been achieved with supervised ML systems and right-censored data in general [6,27,59–62]. The cumulative effect of missing values across several variables and outcomes of interest diminishes the reliability and information density of routinely collected clinical datasets for training and validating accurate ML systems [63–66].

Class imbalance refers to a limitation where one type of outcomes is observed in a dataset more frequently than another. High imbalance is common in medical classification tasks, such as predicting diseases with low prevalence or rare variants of common diseases, monitoring abnormal response to treatment, and preventing clinical equipment failures

[26,47,67,68]. Imbalanced sets also abound in population screening data, where, fortunately, even the most prevalent diseases such as diabetes, only affect a fraction of the population. Training ML classifiers with imbalanced data reduces their sensitivity and thus overall predictive power, unless appropriately accounted for [69–71].

Finally, routinely collected patient data are inherently *noisy* and *prone to errors*. Any locality specific variations, errors or omissions in EMR *aggregate at scale* when combined with population-level datasets. An additional source of "noise" for ML in multimodal and multicentre databases results from inconsistencies in how a given disease symptom is coded, when certain variables are recorded, and how they are interpreted [44,72,73]. For instance, the UK National Health Service (NHS) EMR system accessed in this research utilises over 300 separate codes for direct identification of diabetes mellitus, not including 400 additional product codes related to diabetic medicine prescriptions. Even smoking status identification involves analysis of over 120 read codes, which make distinctions as subtle – and perhaps as prone to arbitrary assignment – as "137*S*.00 Ex-smoker" and "137*K*.00 Stopped smoking". Despite continuous standardisation of clinical databases in the UK and globally, noise, inconsistencies and errors have remained one of the defining limitations of a domain as complex and diverse as human healthcare [47,51,74,75].

Combined, the quality limitations reduce the predictive value of clinical and biomedical datasets. Low information density, high heterogeneity, missing values, class imbalance and errors explain why a seemingly vast multicentre dataset may not contain a sufficient number of observations to effectively approach the clinical problem at hand [53]. Whether using large routinely collected multicentre databases, or single-centre datasets, reductions of already scarce observations often result in datasets as small as *10 observations per predictor variable.*

1.5 Challenges of learning with limited information

Small datasets jeopardise the predictive potential of otherwise powerful ML techniques. Efforts towards data-efficient learning are presently nascent in the ML community, which has been traditionally focused on solving complex problems with Big data [76]. Learning efficiency considerations are emerging in Bayesian optimisation [77–79] and reinforcement learning [80–82], however, the synthetic and real datasets implied in those applications are in the order of tens of megabytes [83] – far beyond what is readily available in many clinical applications, such as hard tissue engineering and organ transplantation.

As a result, ML models trained with insufficiently large datasets often exhibit unstable behaviour in performance, i.e. sporadic fluctuations due to the sensitivity of the ML models to initial parameter values and training order [84–86]. Model initialisation and training algorithms commonly contain deliberate degrees of randomness in order to improve convergence to the global minimum of the associated cost function [14,85,87,88]. With some learning algorithms, the order within which the training data are fed to the model can affect the level of convergence and produce erratic outcomes [85,86]. Moreover, limited test data availability poses a major obstacle to reliably assessing the model generalisation on new samples. Such inter-model volatility limits both the reproducibility of the results and the objective comparison between different NN designs for future optimisation and validation. Previous attempts [89] to resolve the stability problems in NNs demonstrated the success of *k*-fold cross-validation and ensemble methods for a medical *classification* problem; the dataset comprised 53 features and 1355 observations, which corresponds to 25 observations per predictor variable. To the author's best knowledge, effective strategies for classification and regression tasks with less than 10 observations per predictor variable have not yet been

established, thus necessitating the development of a methodology that would enable successful learning from limited information.

1.6 Aims and objectives

The *aim* of this thesis is to develop and validate practical models for clinical outcome prediction and risk stratification based on machine learning with limited biomedical information. Three important clinical applications are addressed by adapting existing, and developing novel, supervised learning techniques for data-efficient regression, classification, and survival modelling.

In hard tissue engineering, the task is to devise a scalable model for hip fracture prediction in severe osteoarthritis based on a small secondary dataset of 35 trabecular bone samples. If successful, the original contribution of this work is two-fold: (1) enabling, for the first time, an accurate and non-invasive estimation of the mechanical strength of a trabecular tissue from a handful of structural and physiological parameters for patients suffering from a severe degenerative bone disease, and (2) evidencing that a small NN model is capable of capturing complex mechanobiological patterns and of making inferences about the diseased bone quality that are inaccessible through mechanistic modelling.

In kidney transplantation, the aim is to reduce the long-term failure risk of donor-recipient antibody-incompatible transplants by providing nephrologists with an accurate and transparent decision support tool. The complexity of developing such a tool from heterogeneous single-centre patient data is that it must combine descriptive and predictive modelling in order to first establish dangerous antibody levels and key risk factors, and then forecast the likelihood of acute and chronic transplant rejection. If successful, this tool for the early detection of acute graft rejection would be the first of

its kind in *antibody-incompatible* kidney transplantation, pioneered in Europe by clinical collaborators from the University Hospitals of Coventry and Warwickshire.

The diabetes screening project, conducted in collaboration with the Nuffield Department of Primary Care Health Sciences (University of Oxford), is aimed at modernising the existing statistical system for managing the early prediction of diabetes in NHS primary health care. The development of a dynamic NN model for 10-year type 2 diabetes risk stratification from 80,000 routinely collected medical records has been stipulated by the study protocol [90]. It has been previously hypothesised that inclusion of blood glucose measurements can increase the prognostic value of the model. To validate this hypothesis, the NN model will be implemented and evaluated in two settings: with and without the blood glucose information.

It is important to note that the ML algorithms developed in this research are not intended to be decision-makers: they are merely statistical tools that allow healthcare professionals to recognise non-trivial high-dimensional patterns that may exist in complex clinical data.

The above applications represent three common clinical tasks: regression, classification and survival analysis, respectively. They also reflect various aspects of limitations inherent in clinical data. The trabecular bone and kidney transplant datasets are limited by size, representing single-centre studies. The diabetes dataset is of a large size, but is grossly incomplete, censored and imbalanced, representing the common attributes of routinely collected data.

The *objectives* of this thesis are listed as follows:

1. to identify effective strategies for managing data quality limitations in the three applications;
2. to develop an application-independent methodological framework for small-data learning (less than 10 observations per predictor variable) and to validate the framework with a sufficiently large external dataset;
3. using existing (1) and novel (2) methodology, to design, implement, optimise and test practical ML prototypes of the healthcare technology required for each application:
    a. an accurate, non-invasive diagnostic tool for depleted femoral compressive strength in osteoarthritic patients of all genders and ages;
    b. an informative, statistically-grounded, and easy-to interpret decision support tool for the prediction, prior to transplantation, of likely transplant outcomes;
    c. a prognostic tool for early indicators of type 2 diabetes in the general population, that would retain high sensitivity, without generating a large number of costly false alarms;
4. to use the clinical insights gained from the ML models in order to detect patients at risk and improve short- and long-term individual outcomes in all three applications.

1.7 Thesis structure

The structure of the thesis is as follows. Chapter 1 introduces machine learning and predictive modelling in healthcare in the context of expert insight generation and clinical data limitations. It outlines the aims and the objective of the thesis. Chapter 2 outlines

machine learning methodologies relied upon in this work, focusing on neural networks, decision trees, and ensemble learning. Chapter 3 presents a novel methodology for the limited data underpinning this research, and describes its validation using a publicly available civil engineering dataset. Chapters 4 and 5 explore the utility of the proposed strategies for data-efficient regression modelling on hard tissue engineering data, and predictive classification on kidney transplantation data, respectively. Chapter 6 describes the challenges of modelling diabetes with large, routinely collected dataset and how they have been addressed with several classification and survival models. The overall contribution and key discoveries are summarised in Chapter 7.

Chapter 2

Methodology

This chapter describes the underlying methodology for the design, training, optimisation and validation of the machine learning models developed for the applications in Chapters 4, 5 and 6, and serves as a foundation for the novel methodology for limited data presented in Chapter 3.

The chapter is organised as follows. Sections 2.1 and 2.2 provide an in-depth explanation of *neural network* and *decision tree-based learning*. Section 2.3 introduces *ensemble learning* and the concept of learner diversity. Section 2.4 presents an overview of the statistical models implemented in Chapters 5 and 6, specifically the *Cox proportional hazards* model and *logistic regression*. Section 2.5 describes *performance validation* and outlines the *criteria* for evaluation of regression, classification, and survival tasks. Section 2.6 details the software and hardware resources utilised in this research. Finally, the sources of primary and secondary data are acknowledged in Section 2.7.

2.1 Neural network learning

2.1.1 Neural network topology and configuration

NNs represent a set of highly interconnected neural computing elements called *perceptrons* (also known as *neurons,* used interchangeably*)* that respond to input stimuli

by crudely imitating the non-linear learning that occurs in biological neurons. In the biological nervous system, an input signal propagates through the *dendrites* to the *cell body*, where a *response* is generated if an excitation (*activation*) threshold is reached, and the response is then fired through the *axon* to the neighbouring neurons. The connections between neurons, called *synapses*, strengthen or weaken depending on how frequently that particular synapse is used to compute a successful response [14,17].

Imitating this property of the synapses, NNs adapt to changes in the environment by varying the strength of individual neural links, referred to as *weights w*, and the inherent inclination of each neuron to produce a predefined output, termed as *bias b*. Figure 2.1 represents a single perceptron that maps an input vector $x = [x^{(1)}\ x^{(2)} \dots x^{(n)}]$ to an output vector $y = [y^{(1)}\ y^{(2)} \dots y^{(n)}]$, each comprising n observations.

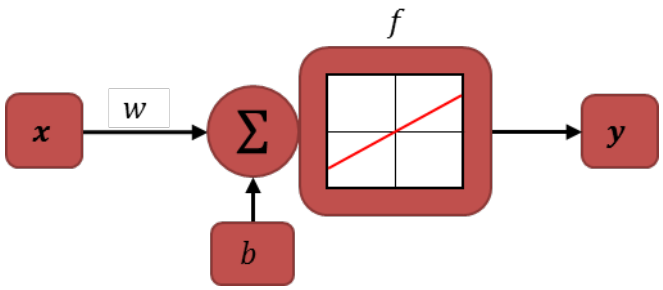


Figure 2.1 Perceptron

The perceptron activation function consists of a summation operation Σ and a transfer function f. Given a linear f, the perceptron in Figure 2.1 computes the output as follows:

$$y = xw + \mathrm{b} \qquad eq.\ 2.1$$

Given p predictor variables, the input becomes a $n \times p$ matrix $X = \begin{bmatrix} x_1^{(1)} & \cdots & x_p^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(n)} & \cdots & x_p^{(n)} \end{bmatrix}$,

where j^{th} column with n observations is one predictor variable $x_j = \left[x_j^{(1)}\ x_j^{(2)} \dots x_j^{(n)}\right]^T$.

Scaling this simple mathematical model of a single perceptron into a multilayer *network* yields a powerful predictive model capable of being trained and learning dynamically from new stimuli in order to map its inputs to outputs [14,91].

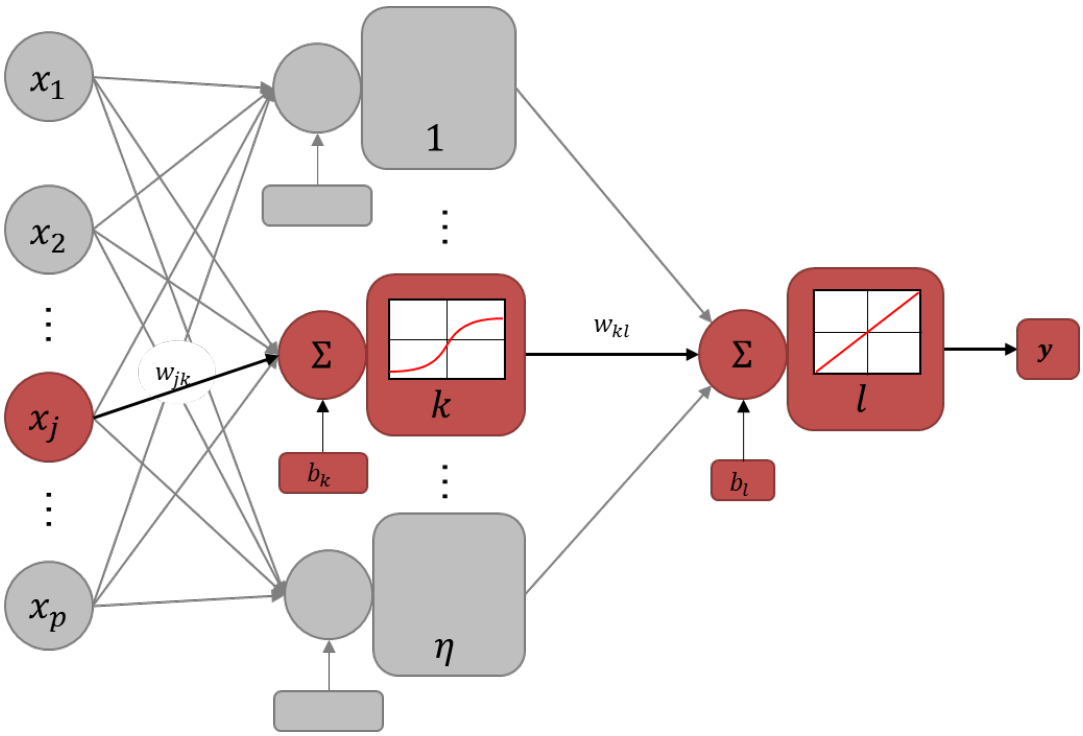


Figure 2.2 Multilayer perceptron network (secondary paths are greyed-out for legibility)

The diagram in Figure 2.2 is an example of a multilayer perceptron network, comprising p inputs, one *hidden* layer with η neurons and one *output* layer with 1 output neuron. The input to the NN does not strictly constitute its own layer, although there is no unanimous consensus in the ML community regarding reference to the separate "input" layer [14,17]. The neurons in the hidden layer connect to every variable $x_1, x_2, \dots, x_p$ in the input X supplied to the network. The *number of hidden layers*, as well as their *size* (i.e. the number of neurons η in the layer) are NN *hyperparameters* that can be tailored for a given application. The number of neurons in the *output layer* is determined by the number of output variables being predicted. In the network depicted in Figure 2.2 the output layer size is 1, since the applications considered in this thesis only required one

output y. The set[1] of biases $b = \{b^{(1)}, b^{(2)}\}$ corresponds to the number of neurons in the network, which in this example is equal to η biases $b^{(1)} = [\, b_1^{(1)}\ b_2^{(1)} \ldots\ b_\eta^{(1)}]$ for the hidden layer and one bias $b^{(2)}$ for the output neuron.

The matrix of *input weights* $W_I = \begin{bmatrix} w_{11} & \cdots & w_{1\eta} \\ \vdots & \ddots & \vdots \\ w_{p1} & \cdots & w_{p\eta} \end{bmatrix}$ comprises elements w_{jk} that connect the j^{th} variable x_j with a k^{th} neuron in the hidden layer. The column vector of *layer weights* $w_L = \begin{bmatrix} w_{1l} & w_{2l} \ldots w_{\eta l} \end{bmatrix}^T$ comprises elements w_{kl} that connect each k^{th} neuron in the hidden layer to the output neuron (in case of multiple outputs W_I is a matrix). There is no mathematical distinction between how w_{jk} and w_{kl} are treated; the input and layer weights comprise a set of network weights $W = \{W_I, w_L\}$. In a *fully-connected* NN every neuron in a given layer is connected to every neuron in the next layer, but not to other neurons in the same layer [17]. The level of connectivity can be customised for a given application, producing more exotic NN topologies such as convolutional layer networks [92], residual networks [93] and Echo state networks [94].

The multilayer perceptron in Figure 2.2 is referred to as a *feedforward* NN, due to its acyclic signal flow in which the signal is propagated from the inputs to outputs and the connections between neurons do not form loops. Other types of flow exist, such as in recurrent NNs [95], auto-encoders [96,97], Hopfield networks [98] and Boltzmann machines [99]. What makes feedforward NNs particularly versatile and effective in the applications considered in this research is that they are capable of approximating arbitrarily closely any continuous function of real valued inputs. This notion of feedforward NNs as *universal approximators* has been elegantly proven by Hornik in 1990 [19].

[1] Braces { } denote sets.

2.1.2 Neural network training with backpropagation

NN training involves determining values for weights W and biases b that reduce the overall network cost function. The *cost function* $cf(y,t)$ refers to the error E between predicted y and target t output value, i.e. $E = cf(y,t)$. In supervised learning, a training dataset $Q = \{X,t\}$, comprising the pairs of input X and target $t = [t^{(1)}\ t^{(2)}\ \dots\ t^{(n)}]$, is supplied to the network. Hence, the aim of training can be defined as *using observations in Q to determine the set of NN parameters $\theta = \{W,b\}$ that minimise $cf(y,t)$*. This task is two-fold: it requires an optimisation algorithm to minimise $cf(y,t)$ and a mechanism to adapt the parameters in θ in response to changes in $cf(y,t)$.

In the NN applications considered in this research, i.e. in Chapters 3, 4 and 6, the task of training NNs is solved by *backpropagation* – a powerful algorithm that has remained dominant in NN development and proved its superiority through time [100–104]. Backpropagation combines the *chain rule* [105], to propagate the slope of E through the network, with an optimisation algorithm, such as *gradient descent*, to compute the necessary changes to network parameters in θ to reduce E. The use of backwards flow through non-linear systems had been well known in control theory for many years before Paul Werbos proposed their application to NNs [101]. The resulting process created two passes of information flow through a network for each *training iteration (epoch)*. In the *forward pass*, an output y was computed from the training sample in the input X. In the *backward pass*, the error E is propagated starting from the output layer through the hidden layers to the input. In so doing, backpropagation updates the values of the weights W and biases b for each neuron, while accounting for their overall contribution to the predicted result. The *forward* and *backward* steps iterate until the error function is minimised, as described in Figure 2.3.

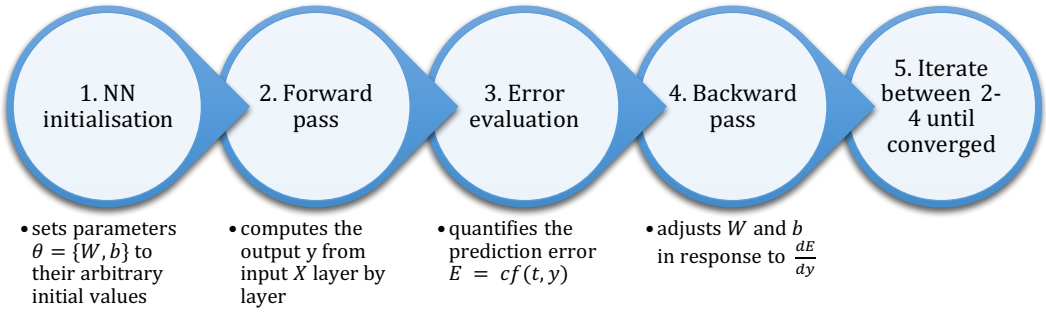


Figure 2.3 Stages of NN training with backpropagation

What constitutes a single *epoch* depends on whether *online* (also referred to as *stochastic*) or batch learning is involved. In online learning, each data sample presented is followed by a weight update, while for batch learning, all data samples from the training set are presented to the NN, and the weight update is calculated for each sample and combined prior to every update event [106]. Online learning requires less computations for each weight update, but it is very sensitive to the outliers, thus making it impractical for the clinical applications considered in this research. Despite being more computationally demanding, batch learning yielded robust performance in the NNs developed in Chapters 3, 4 and 6.

2.1.3 Transfer functions, cost functions and initialisation

Sigmoidal functions, such as $logsig(x) = \frac{1}{1+e^{-x}}$, were particularly suitable for the *hidden* layers of the NNs developed in this research, due to their flexibility in modelling nearly linear, nearly constant and curvilinear functions [17]. Alternative transfer functions are further discussed in Appendix A.1.

The error function $E = cf(y, t)$ is an important part of the backpropagation algorithm, since it determines the dynamics of the learning process. For the *regression* tasks addressed in this research, *mean squared error* (MSE) was particularly well-suited for its

efficiency and ease of interpretability. For a training dataset with n observations in batch learning MSE is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(t_i - y_i)^2 \qquad eq.\ 2.2$$

Since the partial derivative of this squared error function is simply the difference between the target and actual output, i.e. $\delta E/\delta y = t_i - y_i$, this facilitated the efficient computation of weight updates during numerous iterations of NN training.

The *classification* models developed in Chapter 6 used *cross entropy* error, defined as:

$$CE = -\sum_{i=1}^{n}(t_i \ln y_i + (1 - t_i)\ln(1 - y_i)) \qquad eq.\ 2.3$$

The logarithm term $\ln(1 - y_i)$ in $eq.$ 2.3 accounts for the distance between the continuous-valued prediction and the binary target class, providing superior granularity in computing the classification errors. An additional advantage of cross entropy for large datasets, as encountered in Chapter 6, is that the rate at which the NN learns is directly controlled by the magnitude of cross entropy in the output [23,107].

For most error functions, backpropagation is a non-convex optimisation problem, where convergence to a global optimum is not guaranteed; instead, multiple "good enough" local minima are often considered in practice [104,107]. Since NN may converge to a different local minimum for diverse initial conditions, *NN initialisation* plays a crucial role in the optimality of the solution [21,22,107]. The concept of NN initialisation with *random* starting values of W is rooted in the symmetry of the NN topology: if the initial weights are equal across all neurons, then at any given layer all the outputs would be equal too, thus stalling the training process. The magnitude of starting values of θ must

be chosen with care: too small and the training does not proceed, too large and the perceptron transfer function becomes saturated [104]. In order to ensure that the sigmoidal neurons were activated in their linear region, *Nguyen-Widrow initialisation* was implemented [108]. The small positive and negative initial values of θ generated by this algorithm spread the active region of each neuron evenly across the layer input space. The advantage of this approach is that the resulting NN learns the linear part of the X to y mapping first, before embarking on the more difficult, non-linear part [104].

2.1.4 Optimisation algorithms

An *optimisation (training) algorithm* defines the direction and the magnitude of the NN parameter update in response to the derivatives computed in the backward pass [109]. The choice of algorithms defines the behaviour of a given backpropagation network. For the applications considered in this work, two backpropagation algorithms: *Levenberg-Marquardt algorithm* [110] and *conjugate gradient method* [111], proved particularly effective. Understanding the algorithms' behaviour enables appropriate choices in a given NN training task. Since these are standard optimisation functions, their derivations are provided in Appendix A.2 and references therein.

The Levenberg-Marquardt algorithm was used for the regression tasks addressed in Chapters 3 and 4 because it combined the sensitivity of the Gauss-Newton method [112] with the speed of convergence of a simple gradient descent, making it suitable for repeated training with a MSE cost function on small-sized datasets. This was useful when simulating thousands of NNs during the development and validation of the new methodological framework, presented in Chapter 3, without sacrificing accuracy.

The conjugate gradient method, known for its energy minimisation applications in physics [113,114], was an obvious choice for the optimisation of the cross entropy cost

function used by the classification NNs in Chapter 6. A modification introduced by Møller [115] was implemented to the standard conjugate gradient method by scaling the step size, therefore avoiding otherwise computationally-intensive line search. This sufficiently accelerated the algorithm convergence and made *scaled conjugate gradient* (SCG) feasible even for the large-data ensemble NN simulations.

With finite data, determining whether the NN training algorithm has converged is not trivial [116–118]. Instead, practical NNs employ a combination of several *stopping criteria* which have been used in the NNs developed in Chapters 3, 4 and 6. These are:

- minimum gradient (usually <0.00001)
- minimum error (usually <0.00001)
- maximum number of iterations (usually 1000s)
- maximum training duration (pre-defined time in seconds)

Combined, these criteria ensured that the NN training process would terminate eventually: when the cost function ceased to change significantly or the value of the prediction error became negligibly small; or by simply timing out because either the maximum number of iterations was completed or the pre-allocated time expired.

The stopping criteria, however, do not prevent *overfitting,* thus necessitating auxiliary means of controlling the training process, which enable the NN to generalise beyond the training cohort. One way to achieve this is to monitor NN performance on a randomly sampled validation cohort and to stop training early when the validation error ceases to decrease for ω consecutive iterations – a technique aptly named *early stopping*. Alternative methods for preventing over-parametrisation in NNs are *regularisation* (Section 3.5.2) which penalises large weights, and *dropout* [119] which removes the "weakest" neurons.

2.2 Decision tree learning

In the context of ML, *decision trees* (DT) refer to hierarchical learners that map the interrelated consequences (*leaves*) of given decisions (*branches*) based on a predefined reasoning process. The algorithm for classification and regression trees (CART) was first formalised by Breiman et al. [120] and has since been used for descriptive and predictive modelling in medicine [121]. By determining the answer to individual decisions, the tree makes a prediction about a parameter of interest. DTs are nonparametric, i.e. no prior assumptions are made regarding the underlying distribution of the predictor variables [120].

DTs produce a graphical mapping of input conditions to likely outcomes and probabilities, rendering them particularly useful when decisions have to be taken with limited information and reviewed by an expert. The graphical nature of DTs makes them indispensable in clinical applications, where non-technical users might be seeking intuitive and clear-cut representations of the complex relationships in patient data. DTs can be used in medical expert systems for decision making, classification, and probabilistic prediction [121–124].

2.2.1 Nomenclature, topology and configuration

Formally, a *tree* is an acyclic (no loops) directional (top-to-bottom) graph with nodes and edges organised in a hierarchical structure [125]. In a DT, decisions are represented by *branches*, which are arranged in a particular order starting from a *root* node and terminating at a *leaf* node. A data sample traverses through the tree from the root to the leaf following a unique path determined by the decisions at each branch along the way.

A *binary* DT, such as that used for the two-class prediction tasks described in Chapters 5 and 6, separates the input data X_i at a i^{th} *parent* node n_i into two subsets: S_i^{left} and S_i^{right}, so that $X_i = S_i^{left} \cup S_i^{right}$ and $S_i^{left} \cap S_i^{right} = \emptyset$, i.e. the subsets are disjoint (Figure 2.4). The *child* nodes n_{i+1} and n_{i+2} operate on the S_i^{left} and S_i^{right} and divide the dataset into four by producing further child nodes n_{i+3}, n_{i+4}, n_{i+5} and n_{i+6}. When a node cannot be split further, i.e. only one observation remains in its input set S_i or when a pre-defined degree of *purity* is reached, it becomes a terminal leaf node. A node is considered *pure* when all of the observations in that node are of the same target value. The degree of *impurity* is measured by the proportion of observations in the node that do not agree with the majority target value.

The binary split continues until every branch terminates with a leaf node, i.e. all of the observations are assigned to a leaf. The target values could be continued-valued (regression DT) or categorical (classification DT). The binary classification tree model utilised in this work follows Breiman's time-tested CART algorithm [120] as described below.

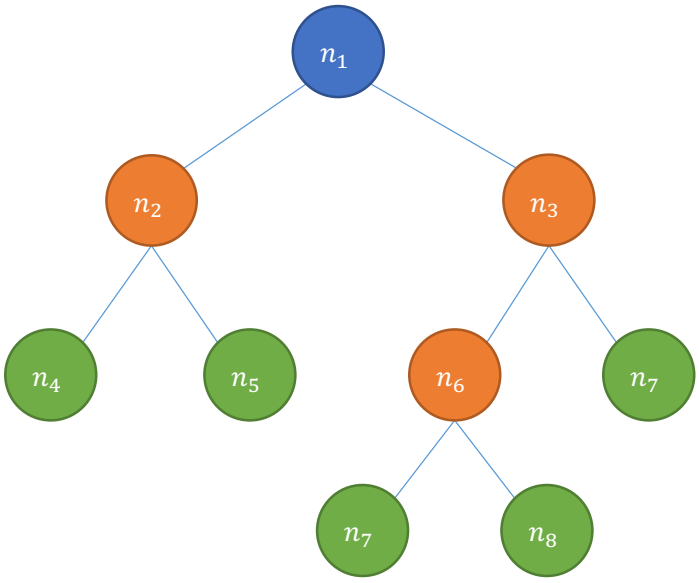


Figure 2.4 A binary DT topology: a root node (blue), three branch nodes (orange), and 5 leaf nodes (green) across 3 levels.

2.2.2 Decision tree training

Training (or *growing*) a DT involves selecting *which* predictor variable is to be considered at each node, and *how* this variable should be split. In order to determine the optimal sequence of decisions that makes a fully-grown DT, the CART algorithm employs *recursive binary partitioning* [17,120]. In recursive binary partitioning an exhaustive search of all possible split values is performed across all potential predictor variables. This greedy algorithm divides the predictor variable range into a number of possible split points, calculates for each candidate split a measure of quality, and chooses the best one to produce two child nodes. Split criteria are discussed further in Section 2.2.3. The process is repeated in a recursive manner for each subsequent child node, until the node is either pure, or when splitting no longer increases predictive accuracy [17]. This top-down iterative process for DT learning is summarised in Figure 2.5. Since not all leaf nodes in a DT are necessarily pure, there exists the notion of *a classification error* – a weighted average of the individual leaf impurities, where weights are the proportions of records in each leaf.

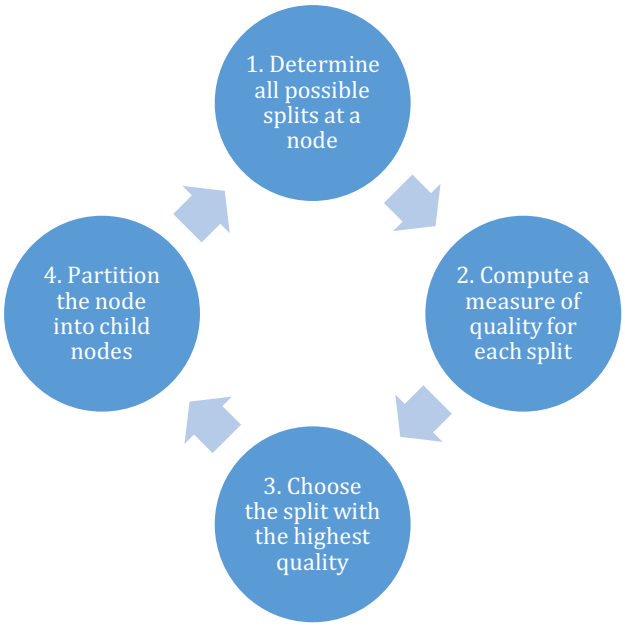


Figure 2.5 Recursive binary partitioning for DT learning

The *number of possible splits* is defined by the data themselves. In this research, both discrete and continuous variables were considered. A continuous variable could be split between any two adjacent values present among the observations, or, alternatively categorised into a smaller number of ranges where appropriate [17]. For categorical variables, all the possible combinations of categories must be considered. The number of these combinations grows exponentially with the degrees of freedom (levels) present: a binary variable offers a single possible split, whilst a variable with l levels could be partitioned in $2^{l-1}-1$ ways. Hence finding an optimal binary split for a continuous predictor is often less computationally intensive than for a categorical predictor with multiple levels.

2.2.3 Split criteria

The *quality* of a candidate split is generally determined by the homogeneity of the target variable in the child nodes it produces [17]. This can be measured in a number of ways. For regression DTs, the most common split criterion is *mean squared error*, which has already been introduced in Section 2.1.3. For classification DTs, the two most common split criteria are *diversity* and *information gain*.

Information gain (or *entropy reduction*) is based on the concept of *entropy* from information theory [126,127]. Entropy (also known as *Shannon's measure of uncertainty*) refers to the average length of the message required to transmit information in variable x. The entropy of x whose k classes have probabilities $p_1, p_2, \dots, p_k$, is:

$$h(x) = -\sum_{c=1}^{k} p_c \log_2(p_c) \qquad eq.\ 2.4$$

Note that the logarithm of base 2 is used because the message length is measured in binary (0 or 1). An obvious problem arises when a node is pure, and thus the probability of one class is zero: would $h(x)$ be undefined? An analogy with signal processing – where information is a *signal* and entropy is equivalent to *noise* – shows that the amount of noise in a crystal-clear signal is zero, as is the entropy of a pure node.

Based on this definition of entropy, the information gain IG for n_i is computed as the difference between the entropy at n_i and the weighted sum of entropy of its child nodes n_{i+1}^{left} and n_{i+1}^{right}, where the weights p_{left} and p_{right} represent the proportion of observations in S_i that reach each child node:

$$IG = h(x_i) - [p_{left}\, h(S_i^{left}) + p_{right} h(S_i^{right})] \qquad eq.\ 2.5$$

It is important to note that information gain is biased towards continuous x and categorical x with multiple levels [120]. The bias correction for information gain could be achieved through a probabilistic p-value criterion with exact randomisation, bootstrapping, Monte Carlo simulation, or asymptotic approximations [128].

Gini's Diversity Index (GDI) is a measure of diversity named after Corrado Gini, adopted from econometrics into machine learning by Breiman [126]. GDI reflects how likely a given observation in the input subset S_i would be misclassified if it was labelled randomly according to the class distribution in S_i. For a classification problem with k classes, the GDI of an i^{th} node with observations S_i is given by:

$$GDI = 1 - \sum_{c=1}^{k} p_c^2 \qquad eq.\ 2.6$$

where p_c is the observed probability of class c samples that reach the node [126,127]. The GDI of a *pure* node, i.e. when the node contains only observations of one class, is

equal to 0. By the same definition, the upper boundary of GDI is always less than 1 for any k, and 0.5 for a k=2 class problem. When applying GDI with recursive binary partitioning, the split that results in a node with the smallest GDI is selected as the optimal split.

Despite the fundamental differences between GDI and information gain, a rigorous analytical comparison by Raileanu et al. found that the two measures disagreed only in 2% of cases [126,127]. The preliminary analyses of the DT models for kidney transplant modelling in Chapter 5 also demonstrated no difference in performance associated with the use of either measures. Given that there was no advantage of using information gain, GDI was used as a split criterion for all models described in Chapters 5 and 6, saving on a small, but recurrent step of computing logarithms.

2.2.4 Controlling leafiness

Just as with NN training, the process of growing DTs was monitored closely to prevent overtraining and avoid unnecessary complexity of a model. In DT, the degree of complexity is determined by the number of branch levels ("depth") and the number of leaves ("leafiness"). Note that if a tree is allowed to grow without restraints, it will achieve 100% accuracy (provided that samples with identical attributes do not have inconsistent class indicators) by fitting every available training sample in X to a separate leaf, producing a "perfect" fit that is bound to exhibit poor performance on additional samples [17]. A drawback of deep trees is that with every new branch node, the subset of samples available for analysis becomes smaller and less representative of the overall performance.

To manage the size of the tree, the following parameters have been constrained in the DT models developed in this research:

- maximum tree depth – the number of branches along the longest path,
- minimum leaf size – smallest permitted observation count per leaf,
- minimum parent size – observation count per node for it to become a parent.

The constraints prevented splits that violated the set limit, and forced the corresponding parent nodes to become leaves earlier than they would otherwise do. Nevertheless, even with a combination of several stopping criteria, DTs may overfit the data. This limitation is inherent in DT learning, since finding a globally optimal tree requires nondeterministic polynomial time, i.e. the problem is *np-complete*. In contrast, practical DT algorithms, including CART, commonly employ heuristic searches, which yield locally optimal decisions at each recursion. Thus, the convergence to a globally optimal tree is not guaranteed.

A more direct, model-based method for controlling overfitting in DT is *pruning*. It is achieved by removing nodes that have the least effect on the overall classification performance [120]. Depending on when the node in question is discarded, two types of pruning exist: post- and pre-pruning.

In *pre-pruning*, the association between the attributes and the target class is assessed by a statistical test (most commonly, χ^2 test [129]), and only statistically significant variables are considered for a candidate split. Although considerably faster, pre-pruning it is less frequently used in practice, as it terminates the learning process prematurely, particularly when the number of observations is small [130].

Post-pruning involves simplification operations such as *subtree replacement* or *subtree raising*. First, the tree is allowed to grow until all observations in the training set are classified correctly. Then the classification error is estimated for the whole tree. Since the error on the *training* set does not constitute a useful estimator, either a separate

hold-out set is required for assessing the pruning error (*reduced error pruning*) or the upper boundary of the confidence interval derived on training error is used (*pessimistic pruning*)[131]. Pessimistic pruning further adds a penalty term to the error at each node, known as *error correction*. As long as the error does not increase, entire subtrees (combinations of connected nodes) may be pruned. In *subtree replacement*, the nodes with the weakest class discrimination are replaced by a leaf representing the majority class in a bottom-up fashion. *Subtree raising*, on the other hand, removes a node with the largest error and redistributes its observations to the next node down in the hierarchy. The resulting tree represents the minimal-complexity model while maintaining the predictive power. For a dataset with p attributes and n training instances, pruning increases the complexity of the DT algorithm from $O(pn \log n)$ to $O(pn(\log n)^2)$. Alternatives to DT learning with pruning, such as decision lists and decision graphs, are discussed in the literature [130].

2.3 Ensemble learning

The accuracy and robustness of NN and DT models can be improved by combining the predictive effort of several learners into a single model, known as an *ensemble* [132,133]. The principle behind a good ensemble is the *diversity* of its constituent models. By generalising over different subsets of an input space, the learners offset mutual errors. The more disagreement there is between the learners in the ensemble, the smaller the overall generalisation error. This relationship between the diversity of an ensemble with k learners and its generalisation performance was explored in detail by Krogh & Vedelsby [56]. Their analytical findings demonstrate that the error for an input sample x in the ensemble $e(x)$ depends on both the individual learner errors $\varepsilon^i(x)$ and the degree of diversity $a(x)$, defined as the variance of the ensemble around the mean:

$$e(x) = \frac{1}{k}\sum_{i}^{k} \varepsilon^{i}(x) - a(x) \qquad eq.\ 2.7$$

Hence, when the learners in the ensemble are strongly correlated, i.e. $a(x) \approx 0$, the ensemble error $e(x)$ would be equal to the average of the errors $\varepsilon^{i}(x)$ of the k individual learners. As the diversity $a(x)$ increases, $e(x)$ decreases. By increasing learner diversity, the ensembles often yield more robust predictive models than any of its constituent learners and offer superior generalisation accuracy [132,133].

2.3.1 Increasing ensemble diversity

The most popular ensembling strategies for increasing learner diversity are *bootstrap aggregation*, *cross-validated committee*, and *boosting*, as described below.

Bootstrap aggregation (or *bagging*) involves training each learner in the ensemble with a different subset of samples drawn randomly from the original training set [134]. Sampled with replacement, each bootstrap subset may contain duplicate observations. All models in a bagged ensemble vote with an equal weight.

To avoid bootstrapping duplicates in bagging, the original sample space could be randomly divided into *disjoint* subsets, producing what is known as a *cross-validated committee* [135]. Similar to k-fold cross validation discussed in Section 2.5, the sample space is partitioned into k disjoint subsets and one subset is withheld. Subsequently, k learners are each trained on k-1 out of the k subsets, with the subset being withheld iterating from learner to learner.

Boosting is similar to bagging, but instead of random sampling, the individual subsets are drawn to emphasise the samples that contributed to the largest error [136]. The process is incremental: each new model is added to target the "weakest" samples of its

predecessor. This is achieved by maintaining a set of weights across all samples and assigning a higher weight to the samples that produced the largest error. Thus, a progressively more difficult problem is being learned by each new model. Boosting allows for improved performance over bagging or cross-validated committees. However, by emphasising a small subset of samples, it is also more prone to overfitting on noisy data [132].

The disadvantage of bagging, boosting and cross-validated committees is that they rely on sub-sampling, which reduces the amount or weight of the training samples available for individual learners [133]. In applications where datasets are already small, further reduction of the sample space may not be feasible. Instead, for the applications with limited data considered in this work, the learner diversity in ensembles was achieved by:

- randomising the initial model parameters (in NN ensembles)
- combining small amount of bagging with random feature sampling (in DT ensembles)

2.3.2 Ensembles of neural networks

The NN ensemble was created by initialising hundreds of NN with random weights and biases. Each constituent NN learner was trained on the complete set. Optiz & Maclin showed that this ensembling approach was "surprisingly effective, often producing results as good as bagging" [132].

Once individual NNs are trained, they can either be merged into a single NN instance, whose weights and biases are a parameter average of the constituent NNs, or they can exist in an ensemble by combining their output predictions [137]. Several approaches

were considered in this research for aggregating the individual outputs $y^{(i)}$ across k learners into a single ensemble output $\bar{y}$:

- simple output averaging: $\bar{y} = \frac{1}{k} \sum_{i=1}^{k} y^{(i)}$
- weighted output averaging: $\bar{y} = \sum_{i=1}^{k} w^{(i)} y^{(i)}$, where $w^{(i)}$is proportional to the "trustworthiness" of the learner i measured by its predictive accuracy
- voting or majority consensus (in classicisation): $\bar{y} = y^{(i)}$ if $\sum_{i=1}^{k} i > k/2$

In the applications addressed by the ensemble NNs in Chapters 3, 4 and 6, simple averaging proved as effective as weighted averaging and voting.

The accuracy of the NN ensemble generally increases with the number of learners until a saturation point is reached, which in turn depends on the amount of noise in the data [118]. Thus, the choice regarding the number of learners k to be included in a given ensemble is a design trade-off between computational efficiency and reproducibility.

2.3.3 Random forest

An ensemble of DTs, aptly named *random forest* (RF), involves growing a number of DTs and aggregating their outputs [138]. The RF algorithms use various degrees of bagging in tandem with *random feature sampling* to further increase the diversity among its constituent DTs. Each tree in the RF is developed with only a subset $p^*$of the total number of input features p, sampled at random. The recommended size of such partial-feature subsets is roughly $p^* = \sqrt{p}$, although empirical results show little sensitivity to this choice of p^*[138]. RFs do not tend to overfit with an increased number of trees, instead, the RF generalisation error reaches a limiting value [138].

By design, the partial-feature trees are unlikely to fit the entire dataset well, hence they are referred to as "weak" learners. Instead, the trees in an RF developed in Chapter 4 were allowed to grow fully and specialise on their specific subset of samples and features, with an expectation that when a large number of them was combined, their performances would benefit from the cumulative effect across the entire sample space [123,125]. This property allowed the RFs to handle numerous input features without having to perform preliminary feature selection or dimensionality reduction.

The built-in feature selection mechanism in RFs also enabled the quantification of relative variable importance. *Variable importance scores* for RF are defined by measuring the increase in prediction error when the values of a variable under question are permuted across the out-of-bag observations; referred to as *permutation test*. These scores were computed for each constituent tree, averaged across the entire ensemble and divided by the standard deviation.

Finally, the predictions made by the individual DTs were combined by *voter consensus*, in which each constituent DT voted for the corresponding class and the majority of votes decided the overall RF output [121,123]. This aggregate vote of several DTs proved inherently less noisy and less susceptible to outliers than a single DT output, and improved the robustness of predictions [132,134,138].

2.4 Statistical methods

The use of machine learning to address the complex research questions considered in this work was driven by pragmatism, not by design preference. Where deemed adequate or where required by existing clinical practice, classical statistical techniques, such as linear, logistic and Cox regression, were initially considered. For example, in Chapter 5 the task of evaluating factors associated with long-term kidney graft survival was

accomplished using the *Cox proportional hazards* survival model, while *logistic regression* was used for discriminating between binary rejection/non-rejection patient groups. In Chapter 6, the Cox proportional hazards model was used to reproduce the existing clinical benchmark model, against which the NN ensembles and survival DT models were subsequently assessed. Classical statistical models continue to play an important role in the exploratory analysis and model benchmarking of modern machine learning algorithms. The two approaches used in Chapters 5 and 6 are described below.

Logistic regression (LR) is a multivariate parametric model that has been widely used in clinical literature due to its ability to infer categorical outcomes [139,140] and thus, address questions such as *Would the recipient reject the transplant?* or *Would this patient be diagnosed with diabetes in 10 years?* LR is a particular case of *generalised linear models* that uses *logit* link function to express the *log-odds* of dichotomous outcome y in terms of probabilities $P(x)$ of p-dimensional input $x = \left[x_1\ x_2\ \dots\ x_p\right]^T$ [141]:

$$\log\frac{P(x)}{1-P(x)} = \beta_0 + \sum_{j=1}^{p}\beta_j x_j \qquad eq.\ 2.8$$

The solution to the $eq.$ 2.8 is presented by the set of parameters β that maximises the log-likelihood LL of the model with 2 classes containing n_1 and n_2 samples:

$$LL = \sum_{i=1,\ y_i=1}^{n_1}\log P\left(x^{(i)}\right) + \sum_{i=1,\ y_i=0}^{n_2}\log\left(1-P\left(x^{(i)}\right)\right) \qquad eq.\ 2.9$$

The value of e^{β_j} (known as the *odds ratio*) explains how the probability of the outcome y being positive changes as variable x_j increases by one unit (if continuous), or by a factor (if categorical). Unlike least-squares models, logistic regression does not stipulate strict assumptions on linearity between x_j and y, nor the normality of x_j, which is

particularly advantageous for modelling medical and biological systems. Nevertheless, when the relationship between x_j and *log odds* of y is linear and X is multivariate normal, logistic regression yields more stable solutions and stronger variable significance [140–142]. When variable associations are unreliable, such as in cases with limited data, a *likelihood ratio test* can be used to assess the *relative* significance of the predictor variables given a model fit [141,143–145]. The likelihood ratio test measured, for every variable x_j in X, the chi-squared χ^2 significance [129] between LL of the full model and the LL of the nested model without that variable. The computation of χ^2 accounted for the *degrees of freedom* in x_j and penalised more complex models [129,143,146].

Cox proportional hazards (Cox PH) regression is a semiparametric multivariate regression model that was used in this research for problems involving time-to-event data [147,148]. The risk of developing an event z, such as transplant rejection or diagnosis with a disease, at time t is expressed by a *hazard* $\lambda(t)$:

$$\lambda(t) = \lambda_o(t)\, e^{\beta x} \qquad eq.\ 2.10$$

where *baseline hazard* $\lambda_o(t)$ corresponds to the overall model hazard when the explanatory variables are absent, i.e. $x = 0$. By accounting for the length of survival period y, Cox PH is able to make inferences on the right-censored data that are common in longitudinal clinical studies, where patients are lost to follow-up. The solution to the Cox PH regression is the set of parameters β that maximises the probability of the observed event z occurring in i^{th} patient ($z_i = 1$) rather than any other k^{th} patient , given by the *log partial likelihood*:

$$LP = \prod_{i:z_i=1}^{n} (\beta x^{(i)} - log \sum_{k:y_k \geq y_i} e^{\beta x^{(k)}}) \qquad eq.\ 2.11$$

The process of modelling with Cox PH and its extensions, including time-varying covariates and non-proportional hazards, are described in detail by Singer & Willett [149] and references therein.

Model selection. With a limited number of observations, both the Cox PH and LR models benefit from a reduction of the number of predictor variables to the most parsimonious set. Since *a priori* knowledge of which variables to include was not available when the exploratory analyses were conducted, the initial Cox PH and LR regressions were fitted on all clinically-relevant variables. Subsequently, using a popular technique of model selection known as *backwards stepwise* elimination [150], the variables that did not contribute to the improvement of the models' log-likelihood (log partial likelihood for Cox PH) were eliminated from the models one-by-one in an automated manner.

The statistical hypothesis tests used in this research include the two-tailed *Fisher exact test* [151] for categorical variables, and the non-parametric *Wilcoxon rank sum* (also known as Mann-Whitney *U*) test [152] for medians of continuous variables. The null hypothesis of no difference between the groups was tested at the 5% significance level.

2.5 Performance evaluation

In order to evaluate a predictive ML model, the validation subset of data must be separate from the training samples. Popular partitioning strategies [107] considered in this thesis include *random sampling*, *k-fold cross-validation,* and *leave-one-out validation*.

With random sampling, as used in early stopping (Section 2.1.4), the validation subset is formed by sampling without replacement. In k-fold cross-validation [56,153], the entire dataset is divided into k disjoint subsets (folds), out of which $k-1$ folds are used for training, and the remaining fold is used for testing. The process is repeated k number of

times, until all folds have been tested. Leave-one-out validation [153] is equivalent to k-fold validation with the fold consisting of a single sample: the model is trained on all but one sample. For a dataset of size n, the process is repeated n times, and each sample is tested exactly once.

The disadvantage of k-fold validation is that it reduces the number of samples available for model training and produces k different models. Although more resourceful with the data, leave-one-out validation is computationally more expensive and more susceptible to outliers than k-fold cross-validation. The advantage of random sampling is that for limited data it allows for a thorough validation if repeated multiple times with different combinations of samples, even when the validation dataset is small [47]. This property of early stopping was integrated into the *multiple runs* method developed in Chapter 3.

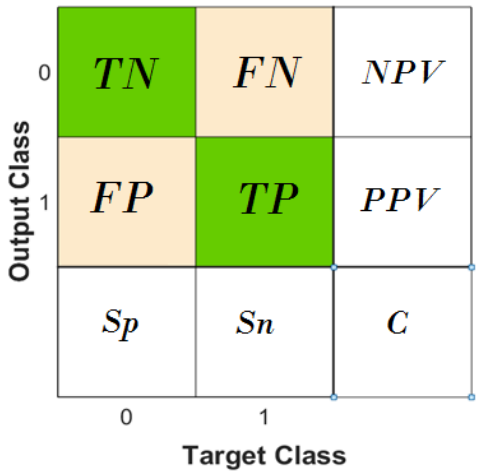
The performance of the predictive models developed in this research was evaluated using the following standard [17] statistical measures:

(i) for regression models: coefficient of determination R and root-mean-square error $RMSE$

(ii) for classifiers: the area under the receiver operating characteristic (ROC) curve AUC, and the measures defined from the confusion matrix (Figure 2.6)

(iii) for survival models: Harrell's C-index and Royston and Sauerbrei's D and R_D^2, in addition to measures in (i).

For classifiers, *confusion matrices* help distinguish True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) observations, and define the classifier's *sensitivity* ($TP\ rate$ also referred to as *recall*), *specificity* ($TN\ rate$), *positive* and *negative predictor values*, and the overall *correct classification rate* (Figure 2.6).



$TP + TN + FP + FN = n$

Correct classification rate $C = (TP + TN)/n$

Sensitivity $Sn = TP/(TP + FN)$

Specificity $Sp = TN/(TN + FP)$

Positive predictive value $PPV = TP/(TP + FP)$

Negative predictive value $NPV = TN/(TN + FN)$

Balanced accuracy $C_{balanced} = (Sn + Sp)/2$

Figure 2.6 Confusion matrix notation and definitions for a binary classifier

The definitions of the remaining criteria in (i-iii) are detailed in Appendix B.

2.6 Software and hardware resources

The applied, problem-driven nature of this machine learning research provided a stimulating environment for the learning of several engineering programming languages, including R and MATLAB™. The code for the preliminary data analysis, neural network design, and the simulation of the new framework in Chapters 3 and 4 was written in MATLAB™ versions R2012b-R2015b for 64-bit Microsoft Windows. The research experiments on tree-based learning in Chapter 5 were implemented in MATLAB™ R2014b. The code for the programming, testing and visualisation of the ensemble NN, Cox PH, LR and survival DT models in Chapter 6 was developed predominantly in the open-source R environment (versions 3.1.1 to 3.4.0), including packages attributed to several authors [154–159].

By most conservative estimates, the number of individual simulations involved in this research amounted to 1,500,000 neural networks and 1600 full and partial decision trees. The computational experiments were conducted on a workstation with 32 GB RAM

and an Intel® Core™ i7-3770 processor with base frequency of 3.40GHz (3.90 GHz achieved with Turbo Boost). Where possible, the simulations were parallelised across the four processor cores. The average simulation time for instantiating, training and logging a run of 2000 small-data backpropagation neural networks was 280 seconds.

The auxiliary programs that supported this research were an open-source Plot Digitizer [160] used for the extraction of data from the literature sources in Chapter 4, and the proprietary DTREG® [161] used to accelerate the computation of likelihood ratio significance tests in Chapter 5. Mendeley Desktop version 1.16.1 [162] was used for referencing and maintaining relevant bibliographical data. For parts of this thesis, Google autonomous speech recognition for English [24,163] was trialled to convert the author's voice to text.

2.7 Sources of data

Four clinical and engineering datasets were used for the training, validation and testing of the ML systems developed in this research. These are as follows:

(1) A civil engineering dataset comprising 1030 samples of concrete from the experiments of Yeh [164] was used for validating the generalising performance of the novel methodological framework developed in Chapter 3. The dataset was obtained through a publicly available Machine Learning Repository at the University of California, Irvine [165].

(2) A hard tissue engineering dataset comprising 35 trabecular bone samples was used in Chapter 4 for the development of a patient-specific hip fracture risk stratification model in severe osteoarthritis. This secondary dataset was extracted through plot digitisation from the original study by Perilli et al. [166].

(3) The kidney transplant data investigated in Chapter 5 were obtained as a result of meticulous examination, recording and follow-up, spanning 14 years, by the UK's leading antibody-incompatible renal transplantation group at the University Hospitals Coventry and Warwickshire (UHCW). The single-centre UHCW dataset containing baseline characteristics and transplantation outcomes for 80 patients was provided directly by the clinical collaborators [167,168].

(4) A UK primary care dataset comprising nearly 80,000 anonymised electronic healthcare records was used in Chapter 6 for the development and validation of novel diabetes risk stratification models. The dataset was obtained through the Clinical Practice Research Datalink in collaboration with a team at University of Oxford Nuffield Department of Primary Care Health Sciences [90].

Chapter 3

Strategies for limited data

As discussed in Chapter 1, limitations on data quality, such as missing values and class imbalance, reduce the size of already small clinical datasets, often below 10 observations per predictor variable. To the best of the author's knowledge, effective ML strategies for such datasets do not presently exist [169–171]. For the three applications considered in this research, existing ML approaches for managing limited data [17,71,172] did not offer a well-rounded solution and, in some cases, were surpassed by biased complete case analysis. In order to bridge this gap, a novel framework specifically for the application of ML to small experimental datasets has been developed in this chapter. This original methodology was pivotal to the successful development of ML models in the three medical applications considered in Chapters 4, 5 and 6.

This chapter commences by describing the strategies for improving incomplete (Section 3.1) and imbalanced (Section 3.2) data used in this research. Sections 3.3 and 3.4 focus on the development of the novel framework for small datasets and its validation with real data. Section 3.5 demonstrates the efficacy of the proposed framework in comparison with existing state-of-the-art techniques.

3.1 Managing incomplete data

Examples of incomplete data abound in medical and biomedical databases. Whether it is a nationwide electronic medical record system or a collection of Excel spreadsheets from a single centre, missing values are one of the defining characteristics of the clinical datasets. Before we can discuss the strategies for handling incomplete data in predictive modelling, it is important to highlight the mechanisms by which the data become missing. As formulated in the seminal work by Rubin [173], depending on whether the missing values are related to the underlying variables, the data are said to be:

- *missing completely at random* (MCAR), if the probability of the data missing does not depend on any variable in the dataset, nor on the response being predicted;
- *missing at random* (MAR), if the probability of the data missing is independent of the response, but may depend on the observed values of other variables in the dataset;
- *missing not at random* (MNAR), if missing data are dependent on the values of the unobserved data.

Whilst with MCAR and MAR the missing data mechanism could be deemed to be independent of the response, MNAR indicates that the missing data may contain information about predicted response [65,172]. Most algorithms for missing data operate under MCAR or MAR assumptions; no effective approaches exist for MNAR data unless the underlying missing data mechanism could also be specified and learned [63,65,172]. Sections 3.1.1-3.1.4 describe common strategies for handling missing data from case deletion to imputation, and model-based induction specific to decision trees.

3.1.1 Complete case analysis

When the missing data mechanism is MCAR, the observations with missing values could be safely omitted from the analysis without introducing bias [65]. Complete case analysis considered in Section 6.3.3 illustrates the potential of small-data NNs when routinely collected patient records contain no missing information across a few variables of interest. Such list-wise deletion is a simple yet appropriate solution for building inferences on MCAR data, but when the data fall under MAR condition or when the proportion of missing values is high, complete case analysis would produce biased estimates of the response [63]. In the case of MAR, the significance of the bias is dependent on the quantity of missing values, and on the degree of association of the missing variable with other confounding variables [63].

3.1.2 Single value imputation

Imputation is required when the discarding of partially incomplete observations is not feasible due to the high historic cost of collection or sensitivity to the MAR bias. Single value imputation aims to substitute the missing values with an estimate. This could be a global or group mean, median (or mode if the value is binary) measured on the observed values of the missing variable, or a value estimated through a model-based algorithm, such as linear or logistic regression, *k*-nearest neighbours or expectation maximisation [63,65].

Substituting missing values in a variable with the *mean, median or mode* of observed values distorts the distributions of that variable; hence the technique is also not suitable for datasets with a high proportion of missing values. A more practical approach, which has been used in Chapter 6, is to complement the imputed variable with an additional

binary indicator flag that identifies which observations in that variable are missing [174]. This strategy is particularly effective with models that can incorporate variable interaction, such as Cox proportional hazards and neural network models discussed in Chapter 6.

Model-based imputation with *linear or logistic regression* treats the incomplete variable as dependent and regresses its values based on all other variables in the data model. The quality of these estimations depends on the quality of the model fit, and is therefore not feasible in applications where a well-fitting model could not be defined in the first place, as is common with physiological models [174–176]. The quality of the estimation could be improved using an iterative e*xpectation maximisation* algorithm [65,172], but it was also deemed inadequate with the level of incompleteness frequently observed in routinely collected population data. For the diabetes data in Chapter 6, expectation maximisation has also proved prohibitively slow, since its rate of convergence increased exponentially with an increasing proportion of missing information.

Nearest neighbour technique was another promising method for handling incomplete data, with some resemblance of how human experts complete missing data [65,177]. For each sample with missing values, the algorithm finds the k most similar samples (neighbours). The missing value is subsequently substituted with the mean value across k neighbours. The drawback of k-nearest neighbour imputation is that with a high proportion of missing values, it alters the data distributions in a way that hinders subsequent response prediction [65,177]. This flaw was observed when the k-nearest neighbour technique was implemented for the diabetes data in Chapter 6 and eliminated during the preliminary data analysis. Nearest neighbour imputation was also considered for the kidney transplant patients analysed in Chapter 5, but it was superseded by the

built-in mechanism for dealing with missing data in decision trees, as described in Section 3.1.4.

3.1.3 Multiple imputation

The problem with single value imputation is that it does not account for uncertainty in the missing values, nor for biased interactions of multiple missing variables [63]. *Multiple imputation* overcomes these issues by leveraging on the predictive distributions of the missing values and creates multiple versions of the dataset [65,172]. In other words, each missing value is replaced by an m-dimensional *vector* of imputed values. This technique is more computationally intensive than single value imputation, but it preserves the variance and uncertainty in the missing values required for realistic modelling of the response [65]. Multiple imputation does not optimise individual sample accuracy, but instead attempts to reproduce the overall resemblance to a complete dataset by generating multiple datasets. Among numerous practical implementations of multiple imputation, *Multiple Imputation with Chained Equations* (MICE) has been shown to be particularly effective for clinical applications [178–180].

For the clinical application described in Chapter 6, MICE was used for imputation of the three continuous variables x_1, x_2, x_3 with missing values. In the initial step, all missing values in x_1, x_2, x_3 were filled at random. Then, x_1 was regressed on x_2, x_3 and all the other variables $x_4, \dots, x_p$ present in the model (including the response variable), and the missing values in x_1 were substituted by simulated draws from its posterior predicted distribution. Using the newly-imputed values of x_1, x_2 was regressed on $x_1, x_3, \dots, x_p$. Similarly, using the imputed values of x_1 and x_2, $x_3$was estimated from the regression on $x_1, x_2, x_4, \dots, x_k$. As suggested by Van Buuren, the cycle was repeated 20 times to refine the estimations [178]. The entire procedure was repeated m =100 iterations in order to

account for the large proportion of missing values (60-70%). MICE resulted in 100 individual datasets, each modelled separately. The parameter estimates and the corresponding standard errors of the 100 individual Cox PH and LR models were combined according to Rubin's rules [173], and the outcomes of the 100 NN models were combined using ensembling approach.

It is important to note that multiple imputation operates under MAR assumption. For MNAR data, inclusion of additional predictors that affect the missing value allows for partial approximation of MNAR to MAR, but only to a certain degree [174]. The potential bias with MNAR data requires careful consideration, since the response variable forms part of the MICE imputation model. In order to ensure the purity of test samples, model derivation and model validation datasets were imputed separately.

3.1.4 Surrogate splits in decision trees

In DTs, the impact of missing variables could be mitigated by means of *surrogate splits* [121,181]. For each primary split where the variable may be missing, "surrogate" substitutions are constructed from other predictor variables that exist in the model. The goal is to find a splitting point with child distributions most closely resembling the primary split. The surrogate splits are then ranked according to misclassification error, and any split that does not perform better than the "go with the majority" rule is ignored. The split with lowest misclassification error is used as the preferred surrogate split, and if neither the primary nor the surrogate variable are available, each subsequent ranked surrogate split has priority. If no surrogate variables are available, the data sample is classified in the majority direction [181]. Surrogate splits allow DTs to handle missing data without imputation. This important advantage of DTs over other ML models was explored in two medical applications considered in Chapters 5 and 6.

3.2 Balancing strategies

The problem of class imbalance is intrinsic to medical datasets, and occurs when one type of outcome is observed more frequently than another, thus forming a majority and minority classes. An imbalance of 1:11, such as that observed among diabetic and non-diabetic patients in Chapter 6, skews a binary classifier to *null accuracy* of 0.917, meaning that 91.7% accuracy could be achieved by simply assigning every observation to the majority class. Significant class imbalance compromises the learning success of a ML classifier, unless adjusted for. The methods used to address class imbalance can be broadly grouped into: 1) cost-sensitive training techniques, and 2) data resampling techniques, including synthetic sample generation.

3.2.1 Cost-sensitive training

Cost-sensitive training prevents overfitting of the majority class instances by adjusting the classifier cost function. An obvious way to implement this is by using a performance metric that is sensitive to the underlying class distributions, such as AUC (see Appendix B) or the weighted harmonic mean of the classifier sensitivity and specificity known as F-score [71,182,183]. Yet, successful algorithms for optimising AUC or F-score directly are scarce for NNs [184,185], and non-existent for DTs. This is due, in part, to the nature of the performance measures: in contrast to MSE or entropy, F-score and AUC are global measures of the true and predicted class agreement, and are not a direct summation of the error in individual observations. Moreover, AUC is non-differentiable, whilst F-score is not concave, thus requiring approximations and rendering their direct optimisation infeasible [186–188]. Finally, utilising a non-separable global cost function for training, meant that the NN models would lose their ability to adapt *incrementally* with every new sample – a key criteria for making medical prognostic models scalable with future data.

A more practical approach for cost-sensitive training was to impose a weights matrix on an existing cost function, so that false *negatives* are penalised more severely than false *positives* [71]. In the prognostic models developed in Chapter 6, the weighted cost matrix approach was considered for classification DT, NN and LR. The weights were determined from the 1:11 ratio of diabetic and non-diabetic patients observed at 10-years in the model derivation cohort, i.e. the cost of a false negative predictions was stipulated to be 11 times higher than the cost of false positives. Such weighted cost matrix rectified the class imbalance issue in the classification DT, although the overall model structure was deemed inappropriate and the classification DT was replaced later in the study with a specialised *survival DT.*

Against the expectations, the cost-penalised NN and LR models overfitted the diabetic outcome patients and produced an overwhelming number of false positives. Poor specificity in models designed to predict a long-term incidence of diabetes, meant a dramatic increase for the NHS screening expenses, rendering such models infeasible. Lowering the cost penalty from 11 to 1 in the increments of 1, did not improve the overall balanced accuracy, rendering the weighted cost matrix approach not suitable for the applications where it is critical to achieve high sensitivity without jeopardising an adequate specificity. This left the author and her collaborators to seek data-resampling techniques for balancing the class representation in the dataset.

3.2.2 Sampling techniques for imbalanced data

Another approach for balancing a dataset is to increase the number of minority class observations: either by resampling with replacement, or by generating synthetic instances [71]. Two state-of-the-art methods of oversampling were considered for the imbalanced dataset in Chapter 6. *Synthetic Minority Oversampling* (SMOTE) generated

new samples from each minority observation and its nearest neighbours by linear interpolation across each input dimension [189]. The method is considered an effective remedy against severely imbalanced data, but has a serious drawback of blending the boundary between majority and minority classes [68,188,190]. An extension to SMOTE called *Adaptive Synthetic Sampling* (ADASYN), in which the density of the new instances is weighted with respect to the class boundaries, was proposed by He et al. [191]. The improvement comes with a price: ADASYN is sensitive to outliers [70,188,191], which makes it less suitable for modelling rare events, such as type 2 Diabetes Mellitus (DM) considered in Chapter 6.

The simplest and often overlooked approach for imbalanced data is *majority undersampling*, whereby majority class observations are removed from the training dataset. An illustrative comparison of SMOTE and majority undersampling is shown in Figure 3.1.

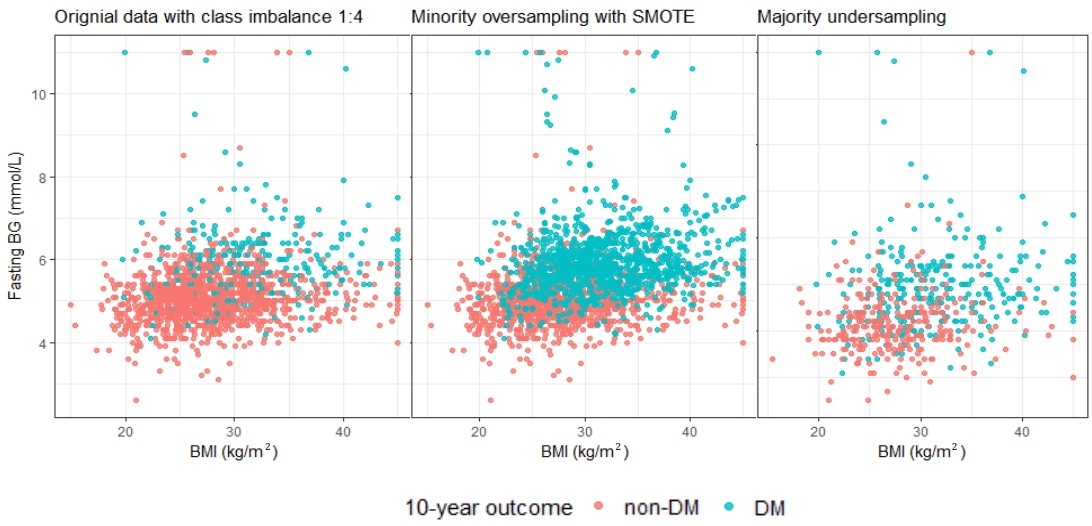


Figure 3.1 Effect of minority oversampling (centre) and majority undersampling (right) on imbalanced data. Two classes of patients correspond to those with known 10-year non-diabetic (DM) outcome (red) and those who had been diagnosed with type 2 DM (teal) during the 10-year. The original dataset (left) comprised 1431 (278 DM) patients with known fasting blood glucose (BG) and Body mass index (BMI).

Although undersampling has been shown to be particularly effective in large datasets with low variance [69,70,188], the omission of a part of the dataset inevitably introduces selection bias. Several techniques have been suggested to reduce this effect by considering majority outliers [190], analysing clusters [68] or cascading an ensemble of classifiers [192]. In this research, majority undersampling was combined with ensemble learning, allowing for all majority samples to be considered at least once, thus effectively removing the inclusion bias. As highlighted in Section 6.3.2, this approach resulted in prognostic models as effective as those built with more complex SMOTE and ADASYN.

3.3 Novel framework for small data

In real world clinical applications, where missing values and class imbalance issues cannot be effectively addressed, the number of samples available for ML training and validation is further reduced. As mentioned in Chapter 1, ML models trained on small datasets (*less than 10 observations per predictor variable*) exhibit sporadic fluctuations in their output, and are difficult to validate. In order to address these issues and improve usability of ML with small clinical datasets, developed in this thesis is a framework consisting of: 1) the *method of multiple runs* for model development, and 2) *surrogate data test* for regression model validation. The method of multiple runs enabled consistent performance comparisons among various ML designs, despite the volatility in predicted outcomes due to small data. Surrogate data test evaluated trained models under small data conditions and provided quantification of the random effects, where additional test samples are not available.

The framework was developed and validated for NNs, although it could be applied to any regression ML system, whose training and initialisation algorithms contain a deliberate degree of randomness. The method of multiple runs, in isolation, is not sensitive to the

nature of the predicted targets, and thus it is applicable to both regression, classification and survival problems. This novel methodology extends beyond the task considered in this chapter and provides a general framework for application of ML to medical problems characterised by limited dataset sizes.

3.3.1 Method of multiple runs

The principle underpinning the *method of multiple runs* is rooted in the simple yet powerful idea of tackling the problem of insufficiently few samples with many independent learners. A large number of NNs of the same design are trained simultaneously. In other words, the performance of a given NN design is assessed not on a single NN instance, but repeatedly on a set (defined here as *run*) of a few thousand NNs. Identical in terms of their topology and neuron functions, NNs within each such run differ due to several sources of randomness deliberately embedded in the initialisation and training routines. For instance, for feed-forward NNs with early stopping these are:

- the initial values of the layer weights and biases,
- the split between the training and validation datasets,
- the order with which the training and validation samples are fed into the NN.

In every run, several thousand NNs with various initial conditions are generated and trained in parallel, producing a range of successful and unsuccessful NNs, as evaluated according to criteria set in Section 3.3.3. Subsequently, the NN performance indicators are reported as collective statistics across the whole run, thus allowing consistent comparisons of performance among runs despite the limited size of the dataset. This helps to quantify the varying effects of design parameters, such as network size and the training duration, during the iterative parameter estimation process. Finally, the highest performing instance of the *optimal NN design* is selected as the *working model*. This

strategy principally differs from NN ensemble methods (as discussed further in Section 3.5) in the sense that only the output of a single best performing NN is ultimately selected as the working model.

In summary, the following terminology applies throughout this chapter:

- *design parameters* are NN size, neuron functions, training functions, etc.
- *individual NN parameters* are weights and biases;
- *optimal NN design* is based on estimation of appropriate hyperparameters;
- *working (optimal) model* is the highest performing instance selected from a run of the optimal NN design.

The choice of the number of NNs per run is influenced by the balance between the required precision of the statistical measures and available computational resources, as larger runs require more memory and time to simulate. In the extreme case of dataset size deficiency considered in Chapter 4, where only 35 samples were available, a consistency to 3 decimal places could be maintained for most performance statistics (such as mean regression between NN targets and predictions) with 2000 NNs, which was deemed sufficient. For inter-run consistency, each 2000 NN run was repeated 10 times, yielding 20000 NNs in total. The average simulation time for instantiating and training a run of 2000 NNs on a modern PC (Intel® Core™ i7-3770 CPU @3.40GHz, 32 GB RAM) was 280 seconds.

3.3.2 Surrogate data test

Where a sufficient number of samples is available, the efficiency of learning of the interrelationships in the data is expected to correlate with its test performance. With small test datasets, however, it is possible for even poorly-designed NNs to achieve, *at*

random, a statistically significant performance. In order to distinguish truly effective NN learners from "lucky" coincidental fittings, it is important to be able to evaluate NN generalising performance in spite of random effects. This is the aim of the proposed *surrogate data test*.

First, surrogate data are *generated* so that they mimic the statistical properties of the original dataset independently for each component of the input vector. Whilst resembling the statistical properties of the original data, the surrogates are *not* meant to retain the intricate interrelationships between the various components of the real dataset. Subsequently, the NNs trained and tested on surrogates are expected to perform poorly. Any seemingly high performance should be deemed coincidental.

The model accuracy on surrogate data informs the NN designer as to what performance could be achieved by a particular NN design due to "luck". By repeating these estimations with multiple surrogate-data NNs, the random effects on the real-data model performance can be quantified. Training and evaluation of multiple NNs with surrogate data is made possible with the method of multiple runs proposed in Section 3.3.1. The highest performing surrogate NN instance defines the lowest performance threshold for real data models. Hence, to pass the surrogate data test, real data NNs must outperform this threshold.

The surrogate samples can be generated using a variety of methods [170,193,194]. For normally distributed data the surrogates were generated from random numbers to match the truncated normal distributions, e.g. mean and standard deviation estimated from the original data, as well as their range and size. For data where individual variables were not normally distributed, random permutations [195] of the original vectors were applied.

3.3.3 Model evaluation and selection

For the regression NNs, the performance of the *individual* models, including the best-performing, was reported using the linear regression coefficients R between the ground truth (targets) and predicted outputs. In particular, regression coefficients were evaluated for the entire dataset (R_{all}), and separately for training (R_{train}), validation (R_{val}), and testing (R_{test}).

The c*ollective* performance of the NNs within a multiple run was reported on the following:

- mean μ and standard deviation σ of R_{test} and R_{all} across all NNs in the run
- the number of NNs that are statistically significant ($R_{all} \geq 0.6$)
- the random effect threshold, $R_{sur,max}$, set by the highest performing surrogate NN, in terms of R_{all} and R_{test}

Under small-data conditions, R_{val} is unreliable on its own for model selection. Hence in the proposed framework, both R_{train} and R_{val} were considered in order to select the best performing model in the multiple run. Although R_{train} does not indicate the NN performance on new samples, it provides a useful estimation of the highest expected NN performance. It is, therefore, stipulated that R_{train} must be generally higher than R_{val} for a well-trained NN. Subsequently, when selecting the best performing NN, the models with $R_{val} > R_{train}$ were disregarded, and from the remaining models the one with the highest R_{val} was chosen. Note that R_{test} should not be involved in the model selection as it reflects the *generalising performance* of NN models on new data.

3.3.4 Summary of the proposed framework

Combined, the method of multiple runs and surrogate data test comprise a framework for application of regression ML models to small datasets, as summarised in Figure 3.2. Multiple runs enable:

i) consistent comparison of ML designs during design parameter estimation,

ii) evaluation of surrogate data and real data models during surrogate data test,

iii) selection of the working model among the models of optimal design.

Surrogate data test provides a mechanism for:

iv) quantification of the random effects due to small data on regression ML performance,

v) validation of the regression ML model performance, where no additional test samples are available.

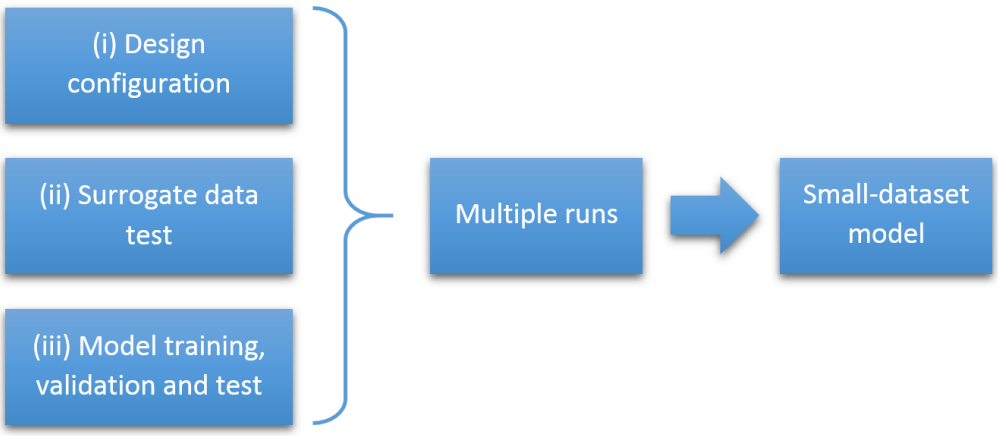


Figure 3.2 A novel framework for the application of regression ML models to small datasets.

3.4 Framework validation

The intended domain of application for the proposed framework was a tissue engineering task of predicting compressive strength (CS) in bones affected by osteoarthritis, as later detailed in Chapter 4. However, the small-data models enabled by the framework could not be considered as valid until it was confirmed that such models were able to generalise on larger datasets. Since large datasets were not easily available in hard tissue engineering, alternative data sources had to be considered. A 1030-sample dataset [164] on CS of another porous solid (concrete) was adapted in this validation study from civil engineering domain.

This dataset enabled, in a principled manner, an investigation of the effects of dataset size on generalising ability of the NNs produced by the framework. Furthermore, access to more data meant that the small-data NNs could be rigorously assessed for generalisation on a large independent test cohort, subsequently confirming whether the multiple runs strategy and surrogate data test were effective.

3.4.1 The concrete compressive strength data

The dataset [164] from 1030 concrete samples was obtained from a publicly available ML repository [165]. It included the following variables (descriptive statistics are provided in Appendix C):

- CS of concrete samples (in MPa);
- quantities of 7 components in the concrete mixture (kg/m^3): cement, blast furnace slag, fly ash, water, superplasticizer, coarse and fine aggregates;
- duration of concrete aging (days).

CS of concrete is a highly nonlinear function of its components and the duration of curing [164], however, an appropriately trained NN could effectively capture that complex relationship between the CS and the other 8 variables. A successful application of NNs to CS prediction based on 700 concrete samples has been demonstrated in an original study by Yeh [164]. The goal of this work was to establish if NNs trained with smaller dataset could achieve comparable performance.

It is important to emphasise that the concrete CS data were used solely for the purpose of validating the proposed modelling methodology, and *not* for transfer learning [11]. Concrete was chosen due to the similarity of the statistical nature of the output (continuous CS) and input parameters, but it obviously had no biological relevance to trabecular bones. In principle, any large dataset with continuous-valued output could have been used.

3.4.2 Effect of dataset size on neural network performance

The dataset on concrete CS was utilised to investigate the role of dataset size on NN performance and generalising ability. It was demonstrated that for a *larger number* of samples the optimal NN parameters could be derived without involving the proposed framework, yet the importance of the framework increases as the data size is reduced.

First, a large-dataset NN model was developed on the complete dataset (1030 samples). The samples were divided at random into training (60%), validation (10%) and testing (30%), i.e. out of 1030 available samples, 630 were used for NN training, 100 for validation and 300 were reserved for testing. Each run comprised 1000 feedforward backpropagation NNs with p=8 inputs and η=10 neurons in the hidden layer. The NNs were trained using the Levenberg-Marquardt backpropagation algorithm [110,196,197]. The cost function was defined as the mean squared error MSE between the output and

actual CS values. Early stopping on an independent validation cohort was implemented in order to avoid NN overtraining and to improve generalisation [153]. The validation subset was sampled at random from the model dataset for each NN, ensuring diversity among the samples. The early stopping criterion ω was set to 10.

Secondly, a NN was applied to a smaller subset of the original dataset. Out of 1030 concrete samples, 100 samples were sampled at random and without replacement [195]. The descriptive statistics of the original and small subsets are provided in Appendix C. The proportions for training, validation and testing subsets, as well as the training and initialisation routines, were analogous to those used for the large concrete dataset NN with an exception to the following adjustments:

i) the size of the run was increased to 2000 NNs to maintain inter-run repeatability

ii) the hidden layer size η was reduced from 10 to 5 neurons to adjust for the smaller dataset

iii) the early stopping criterion ω was reduced from 10 to 6 to reflect the changes in (ii)

Finally, an extreme case with even smaller subset of the data was considered. From the available 1030 samples, 56 were selected at random to yield the same ratio of the number of observations per predictor variable as in the bone CS dataset (35 samples and 5 predictors) considered in Chapter 4. Out of the 56 concrete samples, 41 were used for small-data model development, and the remaining 15 were reserved for model testing. The descriptive statistics on the model and test subset are provided in Appendix C.

Figure 3.3 illustrates the changes to the regression coefficient distributions across a multiple run as the size of the dataset decreased from (a) 1030 to (b) 100, and to (c) 56 samples. All large-data NNs (Figure 3.3 a) performed with statistically significant regression coefficients ($R \geq 0.6$). As expected with large data, the performance was highly accurate, with $\mu(R_{all})$ =0.95 and $\mu(R_{test})$=0.94 when averaged across the multiple run of 1000 NNs.

For smaller dataset NNs (Figure 3.3 b, c), the distributions of the regression coefficients along x-axis were within substantially wider ranges. The standard deviations σ also increased substantially for NN modes based on smaller datasets compared with the initial large-dataset model (Figure 3.3 a). Distributions of the regression coefficients achieved by the 2000 NN instances within the same run (Figure 3.3 c) demonstrate higher intra-run variance when compared to the large-dataset NNs (Figure 3.3 a). Over half of the NNs did not converge and only 762 NNs produced statistically significant predictions.

The mean regression coefficients across the run decreased to $\mu(R_{all})$=0.719, and $\mu(R_{test})$=0.542 (Figure 3.3 c). When considering only statistically significant NNs, the mean performance of all samples was $\mu(R_{all,signif})$=0.839 and individually for tests $\mu(R_{test,signif})$=0.736. Despite higher volatility, an undesirable distribution spread and lower mean performance, the maximal R values for the small-dataset NNs were comparable with those for the large-dataset NNs.

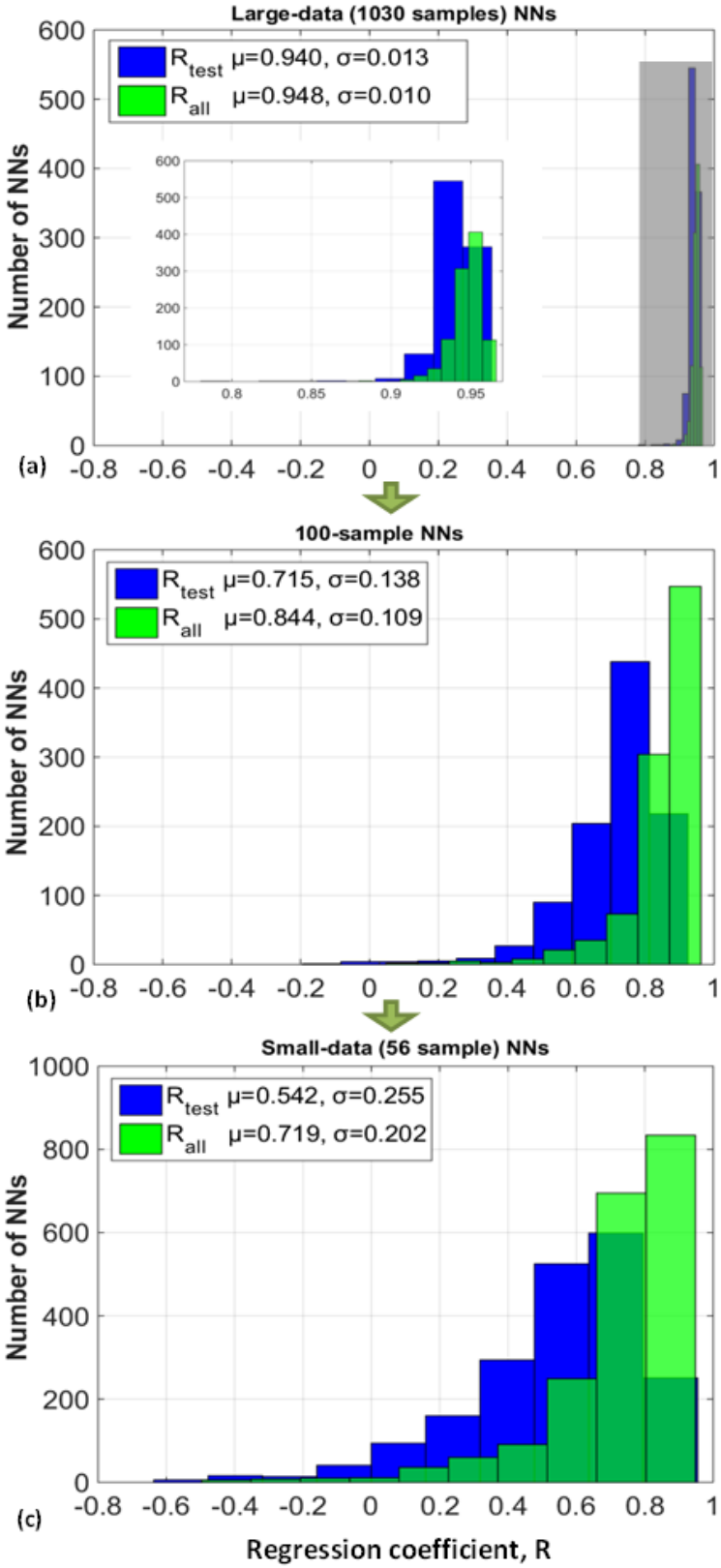


Figure 3.3. Distributions of R_{all} and R_{test} across a run of NNs: (a) large-dataset model (1030 samples), (b) intermediate 100 sample model, and (c) small-dataset model (56 samples). The inset shows the enlarged area highlighted in (a).

3.4.3 Surrogate data test for concrete

As expected, NNs trained on the real concrete data consistently outperformed surrogate NNs. Figure 3.4 demonstrates how the difference in performance between the real and surrogate NNs increased with the dataset size.

For the large-dataset NN developed with 1030 samples (Figure 3.4 a), the surrogate and real-data NN distributions did not overlap. In fact, the surrogate NNs in this instance achieved approximately zero mean performance, which signifies that random effects would not have an impact on NN learning with a dataset of this size.

The 100-sample (Figure 3.4 b) and 56-sample (Figure 3.4 c) surrogate NNs had a non-zero mean performance of $\mu(R_{all,sur,100})$ =0.219 and $\mu(R_{all,sur,56})$=0.187, respectively. They were also characterised by a higher standard deviation of $\sigma = 0.142$ and $\sigma = 0.145$ compared to large-dataset NNs ($\sigma = 0.048$). The non-zero mean performance of NNs suggests that random effects cannot be disregarded with small datasets and require quantification offered by the proposed surrogate data test. Figure 3.4 also demonstrates how the surrogate data test becomes progressively more conservative with decreasing dataset size.

For 56-sample datasets (Figure 3.4 c), the surrogate NNs performed with a mean regression of $\mu(R_{all,sur,56})$=0.187, as opposed to $\mu(R_{all,real,56})$=0.715 for real-data NNs. None of the 2000 surrogate small-dataset NNs achieved a statistically significant performance ($R \geq 0.6$). The surrogate threshold for the 56-sample NN was below statistical significance: the highest performing surrogate NN achieved $R_{sur,max,56}$=0.791, largely due to overtraining, and hence its corresponding performance on test samples of $R_{sur,max,56,test}$ = 0.515 was poor.

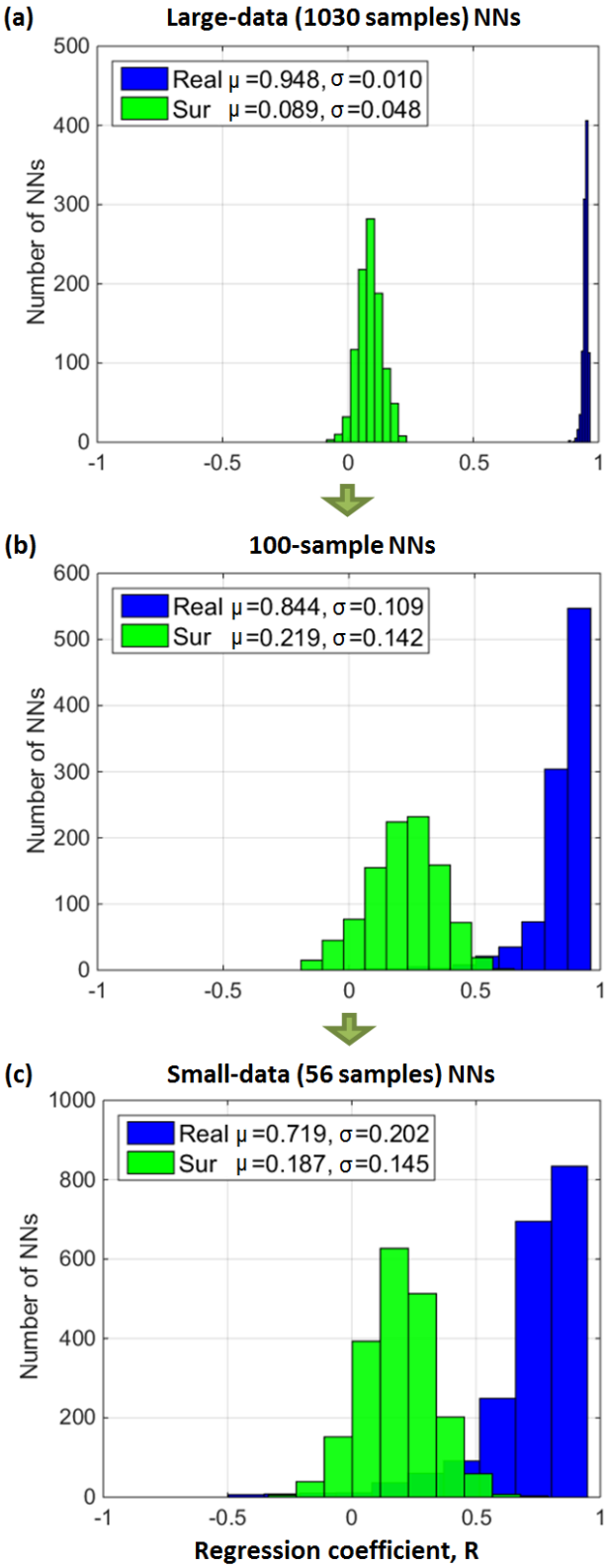


Figure 3.4. Surrogate (green) vs. real concrete data (navy) NN performance for (a) large-dataset model (1030 samples), (b) intermediate 100 sample model, and (c) small-dataset model (56 samples).

3.4.4 Benchmark model

The performance of one of 1000 large-data NN from the run in Section 3.4.2 (Figure 3.3 a) is shown in Figure 3.5. This specimen NN achieved R_{all} =0.944 and generalised with R_{test}=0.94 on 300 independent test samples (Figure 3.5 d). It reflects a benchmark performance of NNs trained with abundant samples using standard techniques.

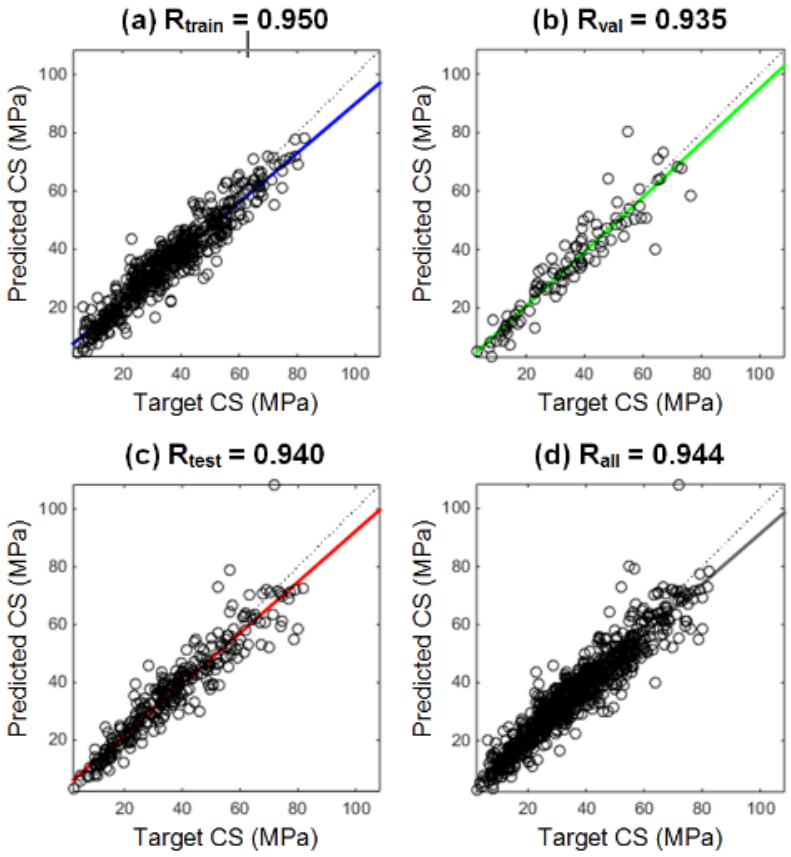


Figure 3.5. Linear regression between target and predicted compressive strength achieved by the specimen large-data (1030 samples) concrete neural network model. Values are reported individually for (a) training (blue), (b) validation (green), (c) testing (red), and (d) the entire dataset (black).

3.4.5 Small-data model developed with multiple runs

Among the 2000 small-dataset (56-sample) NNs, the best-performing NN was selected using the performance criteria defined in Section 3.3.3. This model achieved regression

coefficients of R_{all}=0.92 on the entire dataset, and separately: R_{train}= 0.96, R_{val}= 0.92 and R_{test}=0.90 on 15-sample test (Figure 3.6 a-d). In comparison, the large-dataset NN developed with 1030 samples performed only 2.12% higher. The R values were well above the surrogate threshold $\mu(R_{sur,max,56})$=0.791 determined in Section 3.4.3, hence establishing that high performance of the small-data NN was not due to random effects.

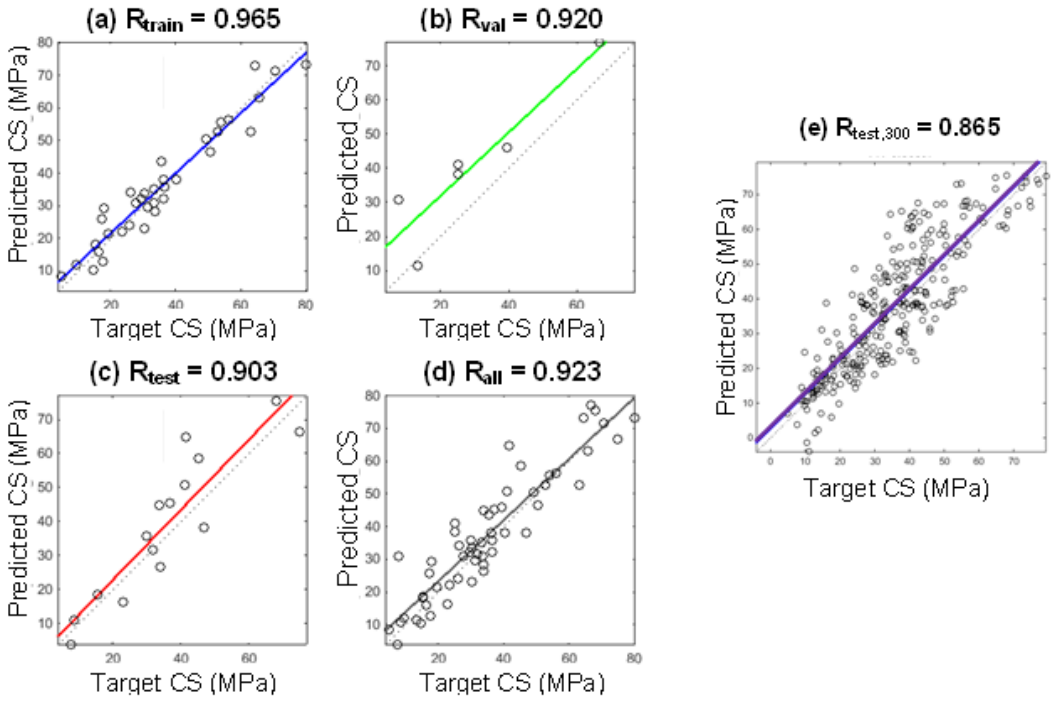


Figure 3.6. Linear regression between target and predicted compressive strength achieved by the small-dataset (56 samples) optimised concrete neural network. Values are reported individually for (a) training (blue), (b) validation (green) and (c) testing (red), (d) the entire dataset (black), and (e) for 300 independent test samples (purple).

To further confirm that the proposed framework was indeed capable of producing well-generalising models from limited data, the performance of the small-data NN was assessed on 300 additional test samples (an equivalent number was used in the large-dataset NN). These were randomly sampled without replacement from the available dataset of $1030 - 56 = 974$ samples not previously seen by this NN. Remarkably, the small-data NN, modelled with only 41 samples, was able to predict CS on 300 new test samples with $R_{test,300}$=0.865 (Figure 3.6 e). The corresponding *RMSE* (as defined in

Appendix B) was 9.5 MPa. This constitutes a 7.5% decrease in generalising performance compared to the benchmark NN, which used 18 (!) times larger dataset (Figure 3.5 c).

In other words, the proposed framework enabled the development *and* validation against random effects of an 86.5% accurate regression ML model on a dataset 18 times smaller than that required to achieve a comparable performance with standard techniques. The remarkable cost-benefit trade-off made possible by the framework highlighted its value for addressing the problems characterised by restricted dataset sizes. The small-data concrete CS NN demonstrated that it was possible for accurate and robust regression ML models to be developed with as few as 56 samples. This finding inspired the clinical applications in Chapters 4-6, which were not previously considered possible with ML due to restricted availability and limited quality of samples.

3.5 Comparison with alternative techniques for small data

Although there are no existing ML techniques for analysis of such extreme cases as the dataset with merely 56 observations, some ML methods, such as NN ensembling, leave-one-out cross-validation, and regularisation, are believed to work well in data-poor situations [198,199]. In order to determine their suitability for datasets as small as 10 observations per predictor variable, it was decided to implement each of the abovementioned techniques and let the practical results speak for themselves.

3.5.1 Ensemble of neural networks

As discussed in Section 2.3, combining predictions of a series of individual NNs into an ensemble often increases their robustness and accuracy [132,133]. In this section, the NN ensemble was compared and contrasted with a single small-data NN model developed with the proposed framework from Section 3.4.5.

The initial NN ensemble was designed by combining the outputs of 1000 NNs trained with the complete dataset (analogous to the large-dataset NNs described in Section 3.4.2 and presented in Figure 3.3 a). As anticipated, this NN ensemble was able to achieve a superior generalisation accuracy of R_{test} = 0.96 when tested on 300 independent samples.

The second NN ensemble was designed by combining the 2000 56-sample NNs (analogous to the small-dataset NNs in Section 3.4.2). This ensemble achieved R_{test} = 0.81 on 15 independent test samples. In comparison, the small-dataset NN model in Section 3.4.5 achieved R_{test} = 0.90 on the same test samples. Subsequently the generalising ability of this ensemble was assessed on 300 additional concrete samples.

The ensemble was able to retain its generalising ability with the accuracy of $R_{test,300}$ = 0.81, proving its robustness, irrespective of the test sample size. Despite such striking consistency between R_{test} and $R_{test,300}$, this generalising performance was found to be 8% lower than that of the single small-data NN developed with the proposed framework ($R_{test,300}$ = 0.87, Section 3.4.5). These results demonstrate that the NN ensemble was able to achieve a remarkable performance on predictive tasks with sufficient data, but was unable to perform as well as the proposed multiple run model on the small dataset considered in this experiment.

3.5.2 Regularisation

The principle behind regularisation is to penalise NN performance for large weights and biases, thus preventing over-parametrisation and resulting in smoother response [107,199]. This is achieved by modifying the NN cost function to consider the sum of squares of the NN weights and biases. Regularisation with *Bayesian regularisation* [200] backpropagation was implemented in the NN models as an alternative technique to early

stopping. The validation samples required for early stopping were now made available for training. Despite this increased training dataset, regularisation did not lead to an increased generalisation performance, and instead it appeared to over-train the NNs (Table 3.1).

Table 3.1 Controlling overfitting with small data: Early stopping vs Bayesian regularisation

Median R across a run of 2000 NNs		
Dataset[2]	**Early stopping**	**Regularisation**
All	0.760	0.864
Training	0.817	0.950
Validation	0.709	
Testing	0.734	0.676

Changing regularisation parameters (such as increasing the *minimum gradient*, *maximum Marquardt adjustment parameter* and *the learning rate decrements*) did not improve the results. Whilst regularisation did not prove superior to early stopping on the extremely small dataset, the strategy independently verified the optimal estimates of the bone CS network hyperparameters, as described in Chapter 4.

3.5.3 K-fold and leave-one-out cross validation

Cross-validation is used to ensure that the results produced by the ML models do not depend on the random choice of the validation set. With multiple runs, the validation cohort used for early stopping was sampled at random from the model dataset for each NN, ensuring diversity among the samples. This validation approach cannot be strictly called a *k-folds cross-validation*: the folds are overlapping and may not cover every possible combination. The resulting validation subsets are more diverse than *k*-folds due to the random sampling in each run of 2000 NNs.

[2] The dataset presented here bone CS data, which is discussed in Chapter 4.

Leave-one-out cross validation was useful for overall inter-run performance evaluation during preliminary design parameter optimisation. However, it was not applicable to the intra-run model selection. Medical applications considered in this research require a stand-alone predictive model, and, therefore, it is advantageous to be able to single out a best-performing NN among those of identical designs. To do so without affecting the purity of test samples, one has to sacrifice a subset for a validation cohort. Whilst being resourceful with an already small dataset, *leave-one-out cross validation* does not provide such independent subset for identification of a single best-performing NN in a run.

3.6 Chapter conclusions

This chapter demonstrated that no general "one-size-fits-all" strategy exists for ML modelling with limited clinical data. The key findings are summarised as follows:

(1) The existing techniques for managing incomplete data rely on missing at random assumption and introduce bias when mechanisms of missing data are dependent on the unknown variables.

(2) Relative disadvantages of list-wise deletion, single value imputation with indicator variable, and multiple imputation with chained equations have to be considered in a practical context.

(3) Decision trees stand out among other ML algorithms for their ability to handle missing data without imputation, rendering these models particularly advantageous for analysing incomplete datasets.

(4) Complexity of the class balancing technique does not necessarily translate into superior performance: majority undersampling combined with ensemble learning was

as effective as state-of-the-art synthetic minority oversampling techniques for data with 1:11 class imbalance.

(5) A novel methodological framework comprising: (1) a multiple runs strategy for predictive model development and optimisation with limited data, and (2) surrogate data test for regression model validation in the absence of substantive test samples, has been developed to address the limitations of small datasets (less than 10 observations per predictor variable) in ML applications.

(6) The framework enabled the successful development and validation of a regression NN with as few as 8 observations per predictor variable and outperformed the ensemble NNs, leave-one-out cross validation and regularisation methods in experiments on small data.

(7) Using the proposed framework, a NN modelled with a small subset of 56 samples was shown to generalise on 300 independent test samples with 86.5% predictive accuracy. This was comparable to the performance of the NNs developed with an 18 times larger dataset using standard techniques, thus demonstrating the remarkable potential of the proposed framework for improving the cost-benefit trade-off in applications restricted by dataset sizes.

Chapter 4

Bone fracture prediction in osteoarthritis

The compressive strength (CS) of a trabecular tissue in the femoral head is indicative of hip fracture risk [201–203]. In patients suffering from severe osteoarthritis (OA), the relationships between the CS and structural parameters of trabecular tissue cannot be explained with the existing mechanistic models [201,204]. In this chapter, a two-layer feedforward NN was developed for modelling CS from porosity, morphology, and the level of trabecular interconnectivity in female and male OA patients of various ages.

Developed with only 35 specimens using the novel methodology for small data proposed in Chapter 3, the NN model was able to accurately (R_{test} =0.983) estimate CS from structural and physiological properties to within 0.85 MPa. Within the limitations of the available dataset, the NN offered a predictive model for clinical and hard tissue engineering decision support. The significance of this work is two-fold: its practical application allows for the non-destructive estimation of strength to femoral fracture in OA patients, whilst also demonstrating the efficacy of the proposed framework for the application of regression NNs to small biomedical datasets.

4.1 Femoral fractures in osteoarthritis

Bone fractures account for more than 20% of orthopaedic hospital cases in the UK, among which fractures of proximal femur (hip) are a growing public health problem

[205–207]. With increasing global incidence, hip fractures are projected to affect 6.26 million people by 2050 [206], which necessitates scalable screening programmes for patients at risk that can adapt to population dynamics.

A fracture occurs when excessive mechanical loading is exerted on areas of the femora during accidental falls and injuries. Such accidents are most common in the elderly as a result of frailty, sensorial and neurological deterioration or muscular atrophy [207]. Whether the femora would fracture at the traumatic impact is determined by the mechanical properties of the femoral tissue at the location of impact. The mechanical properties, in turn, are determined by the quality of the bone tissue, which is depleted with age and hormonal changes in a process known as osteopenia and, in advanced stages, osteoporosis [208]. This is why the residual lifetime risk of hip fracture at 50 years of age is higher in women (20%) than men (5.6%) [206]. Specifically, osteoporosis is characterised by the decrease of *bone mineral density* (BMD), but it is not the only degenerative condition that affects the mechanical strength of the femur.

Osteoarthritis (OA) is a degenerative joint disease associated with the degradation of the articular cartilage and hypertrophic changes in the bone (Figure 4.1) [209]. OA is a life-long condition accompanied by pain and stiffness in the affected joint, loss of dexterity, and reduced mobility, thus considerably limiting the quality of life of the patient. The etiological interplay in OA is complex, but it is recognised that OA in the hip is more common in people with high Body Mass Index (BMI), advancing age (particularly in women), previous joint injury, and genetic predisposition [201,209]. In the UK, 8.75 million people have sought treatment for OA, which translates to the direct annual cost of £5.2 billion to the healthcare system [210]. The treatment of severe hip OA is partial or total hip arthroplasty (replacement).

Whilst it is well-established that osteoporotic decrease in BMD is a key factor in fragility fractures, it might be less known that OA is often associated with *increased* BMD around the joints. Several groups have studied this inverse association between OA and BMD [201,211,212]. Evidence indicates that hip OA modulates the age dependence of BMD in the proximal femora [213]. It was also observed that hip fractures rarely occurred in patients with OA [214], although OA was associated with higher risk fracture in the knee [215] and spine [216]. It is unclear whether the greater BMD in OA could be directly translated into a reduced risk of hip fracture, thus necessitating mechanical modelling of the OA-affected tissue from structural and biological parameters [201].

4.2 Modelling trabecular strength in osteoarthritis

Bone is a living cellular solid with a hierarchical architecture, formed by a load-bearing flexible matrix of collagen and other protein molecules layered with hydroxyapatite nanocrystals [217]. There are two main types of bone architecture: 1) *cortical,* forming the outer layer of long bones with a high density (90% of the volume) of mineralization, low surface to volume ratio and a slow metabolic rate (2.5% annual remodelling), and 2) *trabecular* (cancellous), which is a porous lattice-like structure in the inner bone (Figure 4.1) oriented along stress lines that correspond to maximum load-bearing [208]. It is the mechanical strength of *trabecular tissue* that determines the fragility fracture in proximal human femur. From a structural point of view, the microarchitecture of trabecular bone is characterised by:

- *porosity,* which is measured by the bone volume over total volume (BV/TV) ratio and designates the percentage level of BMD in a given volume;

- *level of interconnectivity*, indicated by trabecular thickness (Tb.Th) and trabecular spacing and which shows how thick trabeculae are and how large the pores are, respectively;
- *morphology,* characterised by the Structure Model Index (SMI) and which provides the 3D measures of the trabecular lattice [218].

These structural properties, together with age and gender related quality factors such as collagen content, mineralisation rate and damage accumulation, define some of the key mechanical indicators of bone fracture risk such as compressive strength (CS), hardness, stiffness and Young's modulus.

It is known from cellular mechanics that CS is related to BMD as a polynomial function (constant exponent of $\frac{3}{25}$) [166,204,219]. This mechanistic relationship explains reasonably well the dependency of CS on BMD in healthy and osteoporotic patients [220,221]. However, for patients with OA, there is an indication that higher BMD does not increase the strength to failure in OA-affected joints [201,211]. Furthermore, the improvement of BMD observed in OA patients may be misrepresented by the heterogeneity of the femoral tissue itself, suggesting that higher bone mineralisation occurs in a cortical bone whilst trabecular bone suffers from osteoporotic *loss* of BMD [204,222]. The task of modelling trabecular tissue becomes even more onerous due to the dependence of CS on local variations in microarchitecture [223]. Computational techniques such as finite-element analysis (FEA) have been used for clinical data in hip fractures with limited success [203,224,225]. All these factors not only highlight the non-intuitive complexity of the physiological and mechanical properties of trabecular tissue in OA, but also emphasise the clinical importance of modelling femoral CS for OA patients, who have been overlooked in the traditionally osteoporosis-centred fracture screening programmes [201,213].

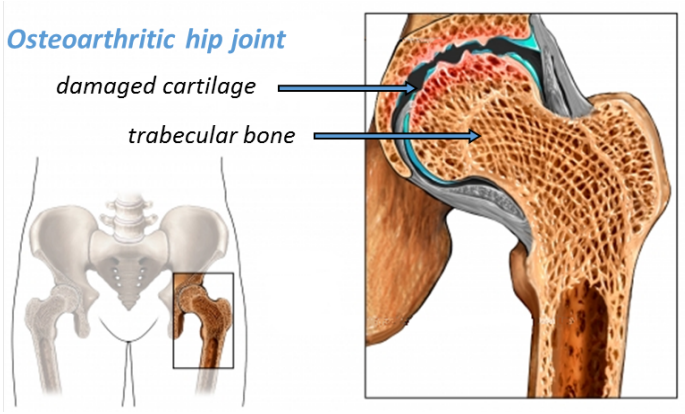


Figure 4.1 Osteoarthritic hip joint

Patient-specific mechanical CS modelling of a trabecular bone is also of interest to hard tissue engineers, who are often faced with the problem of selecting the most successful strategy for both the design and fabrication of synthetic bioscaffolds for the treatment of patients suffering from degenerative orthopaedic diseases triggered by OA, osteoporosis, trauma, injury and metastatic cancer [84,226]. To be effective, bioscaffolds must not only imitate natural trabecular structure for improved bone regeneration, but also match precisely the mechanical loading of the diseased tissue that is being replaced [226,227]. The latter task could be achieved with advanced 3D printing techniques, which are being increasingly adopted in hard tissue engineering [226,227]; provided that the target values of load-bearing CS at the site of implantation could be estimated in the patient femora *prior* to the CS-tailored bioscaffold fabrication.

The clinical application of screening patients at risk of hip fractures in OA requires a predictive, scalable and non-invasive CS model. Firstly, the model must enable the prediction of the dangerously decreased strength to fracture of OA tissue for the timely, preventative intervention in patients of different age and gender groups. Secondly, the model must be scalable with aging population dynamics and the growing incidence of hip fractures. Finally, the model should provide a non-invasive estimation of trabecular CS from the structural parameters available from computer tomography (CT) scans. One

possibility of such modelling was addressed in this research using neural network learning from a small secondary dataset obtained from a published study on the effect of age in OA [166].

4.3 Neural network for bone strength prediction

The use of NNs was motivated by the author's earlier work [84], in which a feedforward backpropagation NN was able to infer, in multi-dimensional space, the complex interdependency between the mechanical and structural parameters of trabecular tissue with a patient's age. The application demonstrated that NNs could cope well with a lack of mechanistic priors and the non-linearity of parameters and determine the significant correlation of CS and age [84], where linear statistical analyses and polynomial BMD-CS models had failed to do so [166].

The remaining obstacle for applying NNs to trabecular CS modelling was the limited availability of training and validation data, since obtaining a labelled dataset with CS values involved the destructive testing of tissue samples extracted through expensive and highly invasive hip replacement procedures. Such complexity is characteristic of the tissue engineering domain, where generation of large-volume and high-quality datasets is highly impractical and often unrealistic. Without effective strategies for small sized datasets, such as the novel framework developed in Chapter 3, NN modelling remained infeasible. The successful application of the framework to concrete samples presented in Chapter 3 demonstrated the small-data potential of regression NNs to CS modelling in porous solids. The NN model for predicting trabecular CS presented in this chapter follows the principles of design optimisation with the method of multiple runs, validation with the surrogate data test and comparison with the ensemble NNs established and detailed in Section 3.3.

4.3.1 The data

This work focuses on a single-centre dataset of structural and mechanical parameters for male and female patients of various ages, adapted from a study on trabecular bones affected by severe OA [166]. Patients affected by secondary OA and other bone and joint diseases were excluded from the original study. Trabecular tissue samples were extracted from the femoral head of 37 patients undergoing total hip arthroplasty due to severe OA. The cylindrically shaped fragment (20 mm in free height and 10 mm in diameter) of trabecular bone was chosen from the principal compressive region of the femoral head and positioned for extraction so that the cylinder axis was aligned with the fixed main trabecular direction for each specimen. Care was taken to ensure consistency of shape, location and alignment.

The physiological data reported were the age and gender of the patients. The structural parameters, comprising trabecular thickness factor Tb.Th (μm), bone volume fraction BV/TV (%), and SMI (dimensionless) were estimated from micro-CT scanning at the isentropic pixel resolution of 19.5 μm with a complete rotation over 185° through voxel analysis and spherical estimation [166]. Tissue strength to compressive failure CS (MPa) was measured from the extensometer ultimate stress readings during deformation testing of the extracted femoral specimen [166]. Finally the tissue samples were subjected to ashing at 650°C for 24 hours with subsequent apparent density measurements in order to confirm the BV/TV values estimated from micro-CT images [166].

Since the apparent density and BV/TV values in Perilli's experiments were the characterisation of the same porosity parameter, they were linearly correlated ($R2 = 0.89$, $p<0.01$ [166]. Whilst essentially conveying the same information about the tissue

as BV/TV measured from micro-CT scans, the measurements of apparent density involved invasive sample collection and ashing, and thus was excluded from the CS model proposed in this study.

The model dataset comprising BV/TV, Tb.Th, SMI, CS and age parameters for each gender was extracted through the digitisation of the nine plots presented in the primary source [166]. The precision error of data extraction was less than 0.7% for any given measurement. CS data were not recorded in one specimen and BT/TV values were missing in the plots for another, reducing the available dataset to 35 samples (17 male and 18 female). The values of the extracted dataset are provided in Appendix D. The dataset was divided at random into training (63%), validation (17%) and testing (20%) subsets, i.e. 22, 6 and 7 samples, respectively. The relative proportions for the testing and validation subsets were lower than in the small-data concrete CS model (Section 3.4.2), since further reduction in the number of samples in the validation set was not feasible.

4.3.2 Small-data neural network design

Considering the size and nature of the available data, a two-layer feedforward backpropagation NN was chosen as the base for the bone CS model, with 5 input features and 1 output (Figure 4.2). The hidden layer neurons implemented a tan-sigmoid transfer function [228], while the output neuron computed the CS output from the input by using a simple linear transfer function.

For every sample, the input vector x contained 5 predictor variables in the following order: x_1 = morphology (SMI), x_2 = level of interconnectivity (Tb.Th), x_3 = porosity (BV/TV), x_4 = age and x_5 = gender. The 5 x η input weights matrix W_I, the η x 1 column vector of layer weights w_L, and sets of biases $b^{(1)}$ and $b^{(2)}$ corresponding to each layer

were initialised according to the Nguyen-Widrow method [108] in order to distribute the active region of each neuron in the layer evenly across the layer's input space.

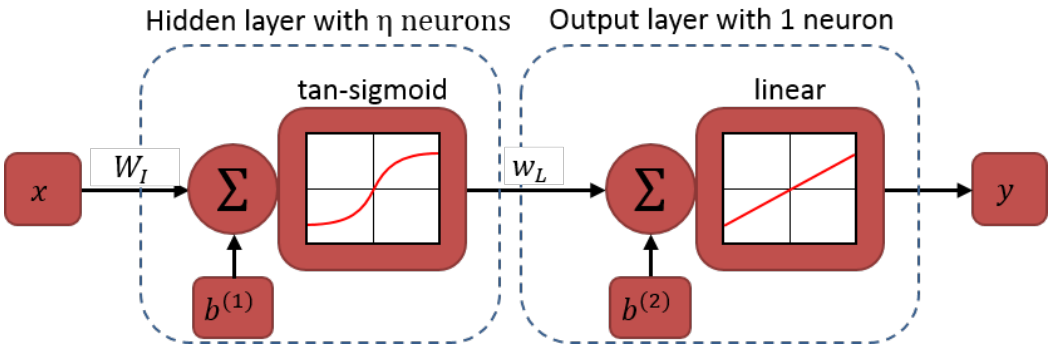


Figure 4.2 The bone CS NN model topology and layer configuration represented by a 5-D input vector, 1 output variable, and one hidden layer of η neurons.

The NNs were trained using the Levenberg-Marquardt backpropagation algorithm [110,196,197]. The cost function was defined by the MSE between the output and actual CS values. Early stopping was implemented in order to avoid NN overtraining and hence ensured better generalisation [153]. The final values of the NN parameters W_I, w_L, $b^{(1)}$ and $b^{(2)}$ were determined during NN training and provided in Section 4.3.4. For every sample with inputs x, the resulting NN model computed the output y (in MPa) as follows:

$$y = tansig\left[xW_I + b^{(1)}\right] w_L + b^{(2)} \qquad eq.\ 4.1$$

During each iteration (epoch), the performance of the NN on training, validation and test samples was monitored in terms of its cost function. Figure 4.3 shows that the prediction error on the training set monotonically decreased with each epoch. The errors on the validation and test samples were sporadic until the 14th epoch. At the 31st epoch the validation error failed to decrease for 9 consecutive iterations and the early stopping criterion was triggered. The weights and biases were then reverted by 9 epochs to the

state at which the validation error was the lowest, i.e. the final state of the trained NN weights and biases corresponded to the 22nd epoch. It is important to note that the 22nd epoch was not the state that minimises cost function for the *test* samples, as these independent test samples were not involved in the model training; their corresponding cost function is provided for illustrative purposes only.

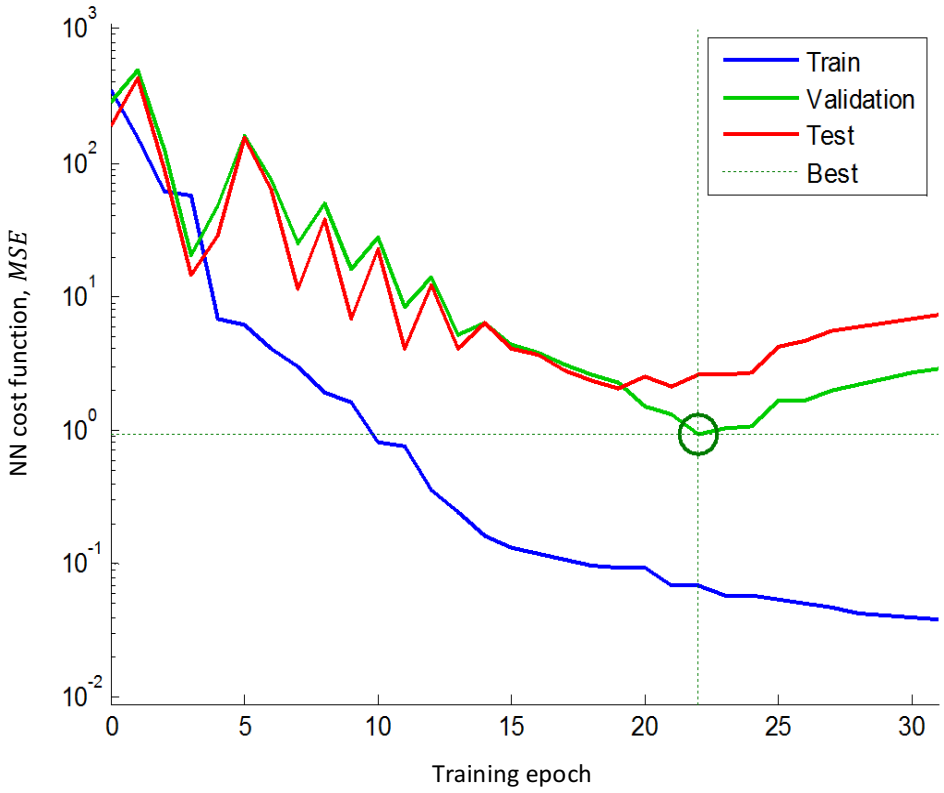


Figure 4.3 NN cost function dynamics during the 31 epochs of training (blue), validation (green) and testing (red). The training was completed upon reaching the minimum validation error (green circle).

4.3.3 Hyperparameter optimisation using multiple runs

The limited availability of training samples stipulated careful selection of the NN design hyperparameters, specifically, the size of the hidden layer η and the early stopping criterion ω, in order to achieve efficient training and improve generalisation. The effect

of increasing the number of neurons η in the hidden layer on NN performance was investigated in a series of experiments involving 20 short runs of 100 NNs.

Reported in Figure 4.4 are the distributions in R_{all} achieved when varying the hidden layer size from η =1 to η = 20 neurons. Despite the inter-run volatility in the results, it was established that the performance was significantly lower in η =1 configuration (p<0.01 for all pairwise comparisons) signifying *under*training. Increasing the number of neurons past η =10 did not improve the NN performance and instead resulted in a gradual decrease in R_{all}, which signified *over*training. Between η =2 and η =10, the highest median R_{all} was observed in η =4 configuration, although it was not statistically different from η =3 (p=0.192) and only marginally different from the η =6 (p=0.036) configurations. The effect of η =4 was also observed on R_{val}. Computed from the leave-one-out validation cohort, it indicated marginal, but statistically significant optimality (p<0.05 for all pairwise comparisons apart from η =3).

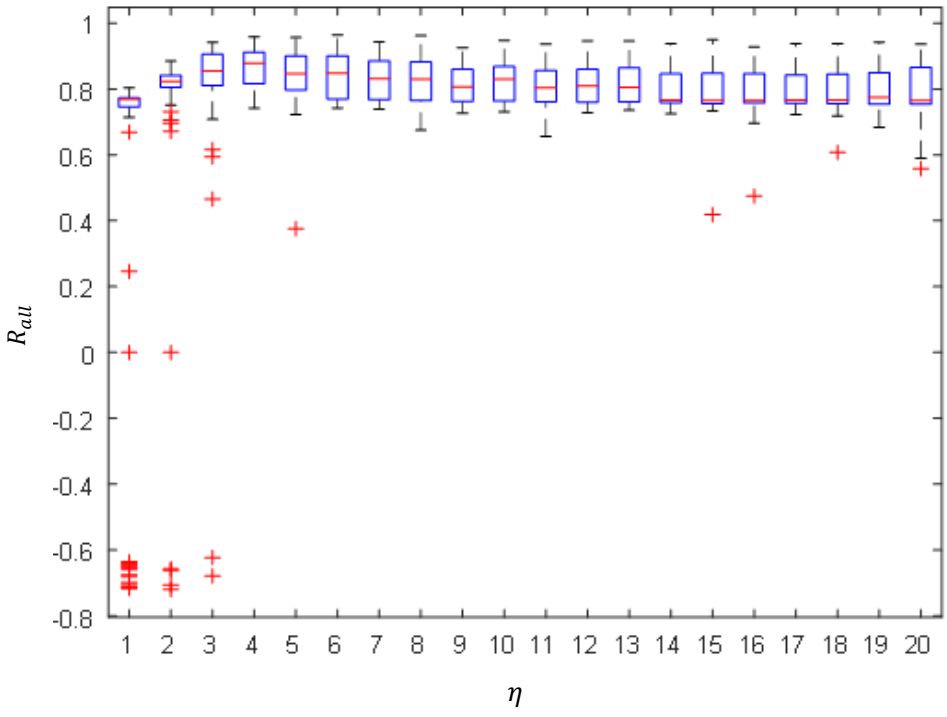


Figure 4.4 Effect of hidden layer size on NN performance

The second hyperparameter considered in the design optimisation was the NN training duration, which is controlled by the early stopping criterion ω. This criterion specifies the maximum for consecutive training epochs the NN validation performance could decrease before early termination of the training process was triggered (Section 2.5).

The effect of ω was investigated on NN performance when varied from 1 to 20 in increments of 1 and from 10 to 100 in increments of 10. When investigated with the total 28 runs of 100 NNs, R_{train} was found to increase substantially for ω values between 1 and 10, and then grow monotonically for each value from 10 to 100, signifying no substantial increase in performance past ω=10. Since large ω directly affected the computational efficiency of the training algorithm (Table 4.1), only ω values between 1 and 10 were further investigated.

Table 4.1 The timing effects of early stopping criterion

ω	Average simulation time (seconds) per run of 2000 NNs on standard PC[3]
5	240
10	280
30	590
100	990

When evaluated on R_{val} at η =4, the early stopping criterion ω had a marginal effect on the performance of the NN and no statistically significant median differences in the range between $1 \leq \omega \leq 10$ (p>0.05 for all pairwise comparisons). In the absence of strong evidence for choosing a specific ω, the value of ω = 9 that gave highest R_{val} was chosen for the final NN.

[3] PC specifications: Intel® Core™ i7-3770 CPU @3.40GHz, 32 GB RAM

As highlighted in Section 3.5.2, Bayesian regularisation [200] was considered as an alternative approach for controlling the training duration without early stopping and determining the size of the NNs. By penalising NNs for large weights, the regularisation reduced some NN weights to near-zero values. The remaining non-zero weights were counted as the number of *effective NN parameters*, irrespective of the theoretical NN size. By varying the number of hidden layer neurons η, it was possible to determine which hidden layer sizes resulted in the highest number of effective parameters.

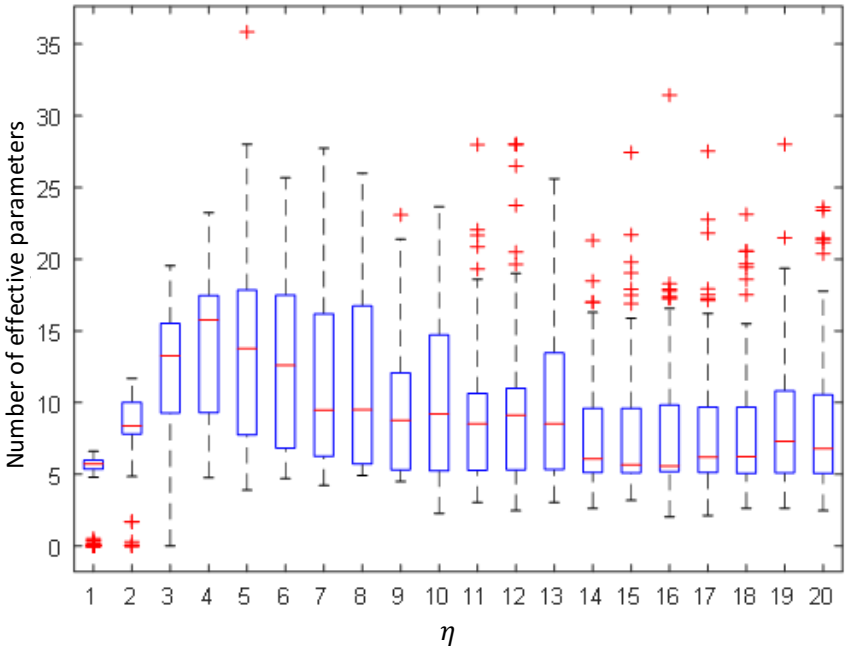


Figure 4.5 Effect of hidden layer size on the number of effective parameters

Reported in Figure 4.5 is the number of effective parameters for a NN configuration with $1 \leq \eta \leq 20$ across 20 runs of 100 NNs. The median number of effective parameters was significantly (p<0.05 for all pairwise comparisons) higher for a η = 4 configuration, indicating that an increase in the hidden layer NN size above 4 neurons introduced redundant near-zero weights, whilst a smaller hidden layer size did not allow maximum NN potential. This observation further confirmed that for the dataset at hand, a NN design with η = 4 neurons in the hidden layer was optimal.

4.3.4 Optimised neural network performance

A full run of 2000 NNs of optimal design was trained and evaluated using the multiple runs strategy described in Section 3.3.1. From the 2000 NNs considered, the best-performing NN was selected using the criteria detailed in Section 3.3.3. The resulting NN model was capable of predicting trabecular tissue CS with *RMSE* = 0.85 MPa on test samples.

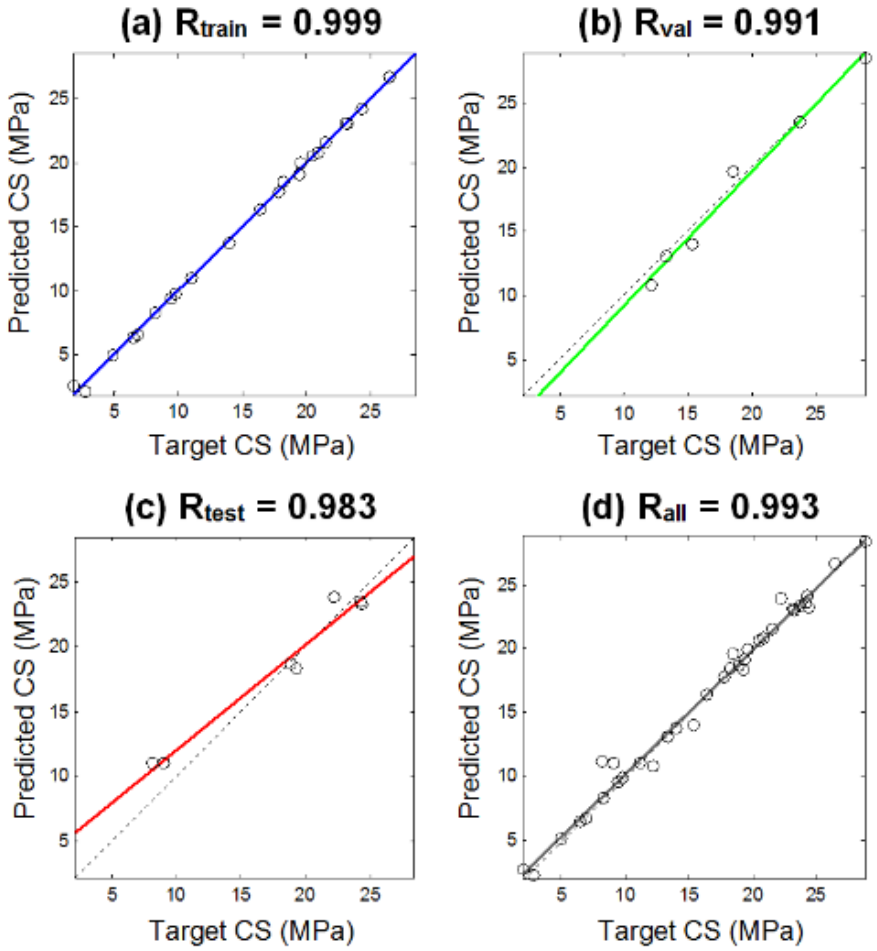


Figure 4.6 Linear regression between target and predicted CS achieved by the small-data bone NN. Values were reported individually for a) training (blue), b) validation (green) and c) testing (red), and d) the entire dataset (black).

The linear regression coefficients between target and prediction achieved by the NN were individually: R_{train}=0.999, R_{val}=0.991, R_{test}=0.983 and in total: R_{all}=0.993 (Figure 4.6 a-d). This indicated a high accuracy of predictions despite the limited dataset (35

samples). The final values of the weights and biases of this fully-trained network complete the unknown terms in *eq.* 4.1:

$$W_I = \begin{bmatrix} 0.887 & 2.382 & -0.888 & -3.584 \\ 1.301 & -1.586 & 0.904 & -3.841 \\ -3.268 & 0.632 & -1.342 & -0.144 \\ -1.216 & -2.153 & -1.380 & -3.000 \\ -0.620 & 1.592 & -0.379 & -1.169 \end{bmatrix} \qquad w_L = \begin{bmatrix} -0.698 \\ -0.151 \\ 2.349 \\ -1.501 \end{bmatrix}$$

$$b^{(1)} = [0.268 \quad -0.006 \quad -1.224 \quad -4.972] \qquad b^{(2)} = 0.623$$

4.3.5 Surrogate data test

The surrogate data test proposed in Section 3.3.2 was used to validate the NNs trained with real data against those trained on surrogate data and, therefore, establish the minimal performance threshold that the candidate real data models must exceed. The surrogates were generated using random sampling to mimic the distributions of the original bone data (Section 3.3.2). The resulting surrogate dataset is provided in Appendix D. When analysed across the total of 20000 NNs in 10 runs of 2000 NNs, the real dataset NNs consistently outperformed the surrogate NNs with, on average, a 35% performance increase (Figure 4.7 a).

The median $R_{all,sur}$= 0.38 and median $R_{all,real}$=0.78 across 20000 NNs were significantly different among the real and surrogate data NNs (p = 0, Wilcoxon rank sum test, Figure 4.7 b). Similar differences in the distributions of $R_{test,real}$ and $R_{test,sur}$ were observed for the NN performance on test samples (Figure 4.7 c-d). The surrogate threshold for the bone dataset was found to be around $R_{sur,max}$ = 0.87. By quantifying the random effects in training and initialisation of the bone CS NNs, the surrogate data test validated that the performance of the real data models above the surrogate threshold was not due to noise.

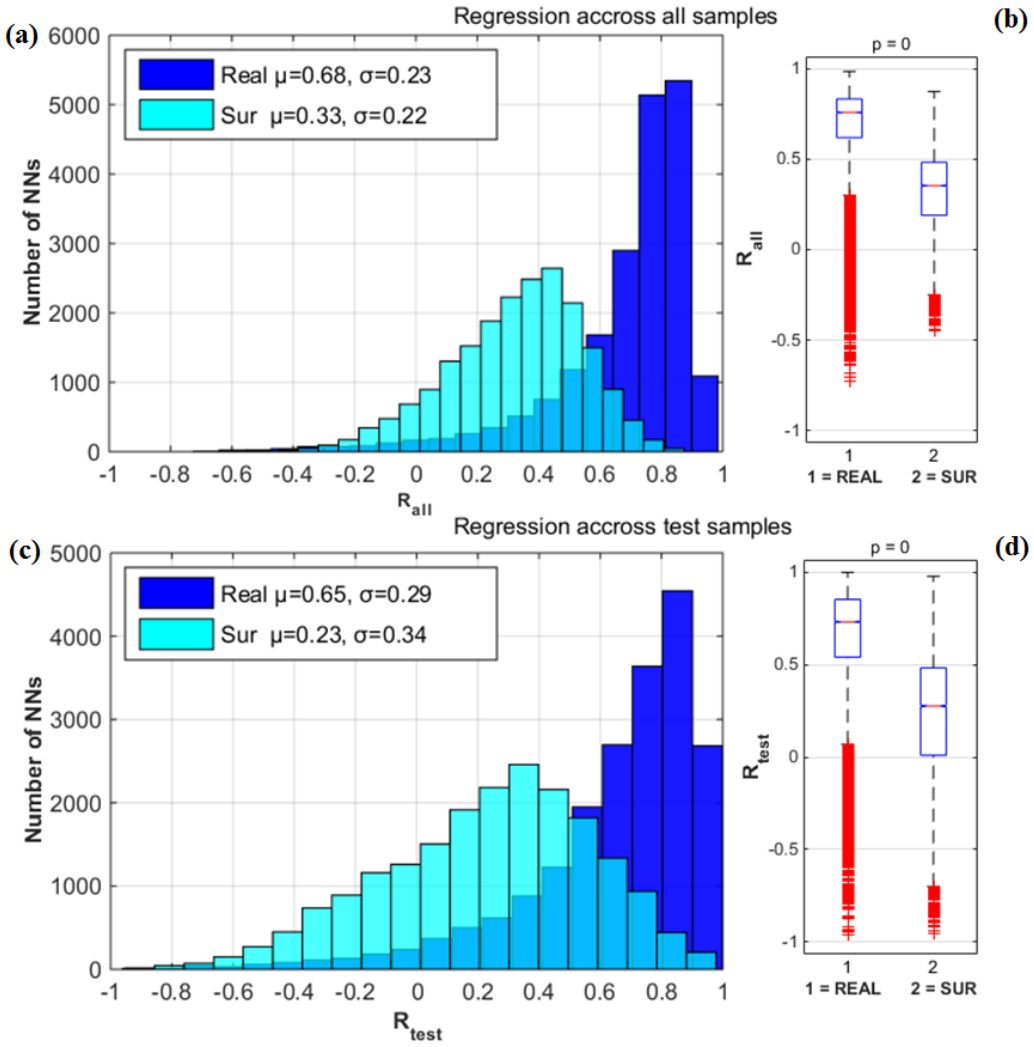


Figure 4.7 Distributions (a) of regression coefficients achieved by NNs for surrogates (light blue) and real bone data (navy) and (b) Wilcoxon rank sum test for medians across all samples. Distributions and Wilcoxon rank sum test results across test samples are reported in (c) and (d).

4.3.6 Comparison with a neural network ensemble

Ensemble learning was implemented by combining 2000 small-data NNs, among which learner diversity was achieved through randomising the initial model parameters (Section 3.5.1) and aggregated using performance averaging (Section 2.3.2). The NN ensemble achieved R_{test} = 0.882, which was 11% lower than the accuracy of the proposed multiple run NN model (R_{test} = 0.983) and only marginally higher than the surrogate threshold $R_{sur,max}$ = 0.87 established in Section 4.3.5 for the bone dataset.

This result further confirmed that NN ensembles, when tasked with small-dataset applications, were unable to realise their full predictive potential and were inferior to NNs designed within the multiple runs framework. One possibility for improving ensemble diversity was to train constituent NNs with different versions of the original dataset, for instance, by resampling or repeating the source plot digitisation. Such an approach was investigated in Chapter 6, whereby each constituent NN was trained on a different version of the imputation dataset.

4.4 Clinical significance and limitations

The application of NNs for hard tissue modelling in degenerative conditions was a novel and largely unexplored area. Among the limited number of relevant studies on trabecular bone modelling, only a few adopted NN-based approaches [229,230]. Habli [230] used a NN model for the estimation of apparent fatigue damage accumulation due to cyclic loading in a trabecular bone from FEA simulations. Zadpoor et al. [229] used NNs to analyse FEA data and model the mechanical loading effects from the spatial distribution of density in the femur. The key limitations of both these studies were the dependence of NN performance on the validity of the underlying FEA model's assumptions, thus necessitating a stand-alone NN that could integrate the complex structural and physiological parameters directly into a single model of a human femur. At the time of publication [84,231–233], the NN model developed in this research was the only known application of such NNs for trabecular tissue modelling in severe OA.

The NN model offered 98.3% accurate predictions of the strength to failure of osteoarthritic hip joints from the structural and physiological parameters of the femoral trabecular tissue in OA patients. In the absence of a comparable CS model specifically for OA, the power model from cellular mechanics, with $R_{power\ model}$ = 0.916, was the best

existing fit to the data [166]. By inferring non-linear variable interrelations in the heterogeneous multi-dimensional dataset, the NN improved $R_{power\ model}$ by over 8% on all samples R_{all} = 0.993.

The high accuracy of the proposed CS model enabled the early stratification of bone fracture risk based on structural and physiological parameters that can be derived without invasive tests on the patient. Hence, by predicting how CS correlates with the bone volume fraction, trabecular thickness and structure model index for OA patients of various age and gender groups, the NN model provided a decision support tool for hard tissue engineers and clinicians alike [234]. The potential practical applications include: the estimation of bone fracture risk in OA patients from CT-scans and basic physiological data, the load modelling of synthetic bioscaffolds that mimic natural trabecular bone damaged by osteoarthritis, and the tailoring of bioscaffold designs for an individual patient to match the damaged trabecular tissue at the site of implantation.

The predictive NN model can be adapted to larger datasets, extended to other degenerative bone disorders, or scaled for modelling new anatomical locations, with a marginal increase in design effort and cost [14,26]. Such scalability is inherent in the underlying ML nature, which enables NNs to learn and improve their performance with new data [36,88,235,236]. Using the proposed NN model for OA as a prototype, future predictive NNs could provide valuable clinical insights for the early detection of patients at risk of hip fractures and for the preventive treatment of bone disorders, thus reducing fractures and improving surgical effects.

4.5 Chapter conclusions

The key findings demonstrated in this chapter are as follows:

(1) The methodological framework for small datasets developed in Chapter 3 was effective in enabling NN-learning from 35 osteoarthritic specimens with an aim to predict trabecular strength to hip fractures from structural and physiological parameters.

(2) The regression NN developed and optimised using multiple runs achieved 98.3% accurate predictions on independent test samples. Further validation with a surrogate data test confirmed that the accuracy achieved by the NN was above the threshold of $R_{sur,max}$=0.87 attributable to random effects due to small datasets.

(3) The NN offered an accurate and scalable predictive tool for the non-destructive estimation of femoral compressive strength in patients suffering from severe osteoarthritis, with potential extension to other degenerative bone and joint disorders.

(4) The proposed methodology confirmed that the size of datasets does not necessarily limit the utility of NNs in the clinical and hard tissue engineering domains.

Chapter 5

Outcome prediction in antibody-incompatible kidney transplantation

Human leukocyte antigen (HLA) sensitisation is a major public health problem that limits access to kidney transplantation for as many as 25%-47% of the patients awaiting a deceased donor transplant [167,237]. The growing field of antibody-incompatible transplantation demands novel insights into the complex association between baseline clinical and immunological indicators and patient outcomes [167,238,239].

The descriptive and predictive models developed in this chapter establish the association of the dominant HLA isotype and its subclasses with both short- and medium-term renal transplant outcomes. A time-to-event graft survival was modelled with Cox PH (Section 5.4.1), whilst acute graft rejection within 30 days post-transplant was explored using logistic regression (Section 5.4.2). A granular and accurate predictive model for early (acute) antibody-mediated transplant rejection was developed with a decision tree classifier (Section 5.5) using the multiple runs strategy for a small, single-centre dataset. This work demonstrated the potential for classification from small clinical data (Section 5.6) and offered novel clinical insights into the area of antibody-incompatible transplantation (Section 5.8).

5.1 Antibody-incompatible kidney transplantation

Organ and tissue transplantation is recognised as an effective treatment for many renal (kidney) pathologies including end-stage renal disease. Transplantation from living or deceased donors can dramatically improve the recipients' quality of life, often offering the only solution for their survival [240]. In the UK alone, over 3100 life-saving and life-transforming kidney transplantations were performed in the past year [237].

For a successful transplantation outcome, the recipient and donor should be matched for tissue proteins called *human leukocyte antigen* (HLA). HLA mismatches between the transplant recipient and their donor may cause the development of antibodies against HLA, which can lead to transplant (graft) failure and endanger the option of a subsequent future transplant. HLA antibodies can also be stimulated by pregnancy and blood transfusion. Patients with *preformed* HLA donor-specific antibodies (DSAs) have longer waiting times for surgery or are unable to receive a renal transplant. Current NHS Blood and Transplant data indicate that among the 5233 patients on the kidney transplant register in March 2017, 32% had been waiting for a suitable graft for over 3 years, with a median waiting time of 864 days [237]. Between 25% - 47% of patients on the deceased donor kidney programme in the UK are unable to receive a transplant due to HLA sensitisation [167].

Antibody-incompatible transplantation (AIT) was pioneered in Europe by the University Hospitals Coventry and Warwickshire [241,242] to enable transplantation procedures on HLA senstised patients. In the past year alone, HLA-incompatible transplantations saved and improved the lives of 654 patients in the UK alone [237]. AIT is becoming increasingly feasible due to the advances in immunosuppressive drugs and surgical techniques that allow for the recipient's DSA levels to be decreased prior to

transplantation [243–245]. Nevertheless, the complete elimination of DSAs and immunological memory is not practical, hence AIT is considered a high-risk intervention: about 40% of HLA-incompatible kidney transplants experience an episode of rejection, which, in its chronic form, leads to transplant failure [237]. The ability of nephrologists to identify patients at high risk of transplant rejeciton *prior* to transplantation is diminished, because neither specific *types* of harmful HLA DSAs nor their *acceptable levels* have been established.

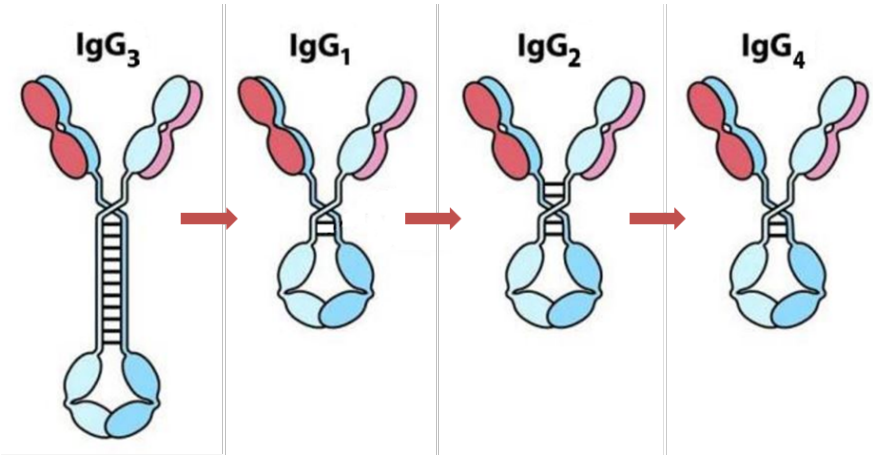


Figure 5.1 Immunoglobulin G molecule structure and class switching. Adapted from [246].

Among isotypes of HLA DSAs, *Immunoglobulin G* (IgG) and its four subclasses, IgG_1-IgG_4 (Figure 5.1), are recognised as principal agents for humoral (antibody-mediated) rejection [247–249]. The four subclasses of IgG exhibit structural and functional differences that may be associated with diverse clinical outcomes [246]. In the small number of studies that have investigated HLA-specific IgG subclass associations with transplant outcomes, some report that IgG_1 subclass DSAs were dominant in poor graft survival [250] and rejection [168], whilst others report the harmful effects of IgG_4 subclass DSAs [167,247]. Predicting AIT outcome from IgG subclass information is further complicated by the *class switching* of IgG_3 to IgG_1 to IgG_2 to IgG_4 (Figure 5.1) – a common phenomenon which occurs as the recipient's immune system develops a humoral response to the transplant [246].

Determining which IgG subclasses are particularly dangerous and establishing their safe levels prior to the surgery could prevent graft loss and/or excessive treatment by harmful and expensive immunonosuppresive drugs. Hence the field of AIT requires both descriptive and predictive models that could leverage multidimensional associations among patient and antibody characteristics with the likely transplant outcomes.

5.2 Machine learning in kidney transplantation

In standard (non-AIT) kidney transplantation, the task of outcome prediction has been considered in a number of studies using machine learning. In particular, decision trees have been a popular choice, likely owing to their graphical interpretability and ability to supplement nephrologists' intuitive insights with data-driven statistical evidence.

Greco et al. studied long-term kidney graft survival and concluded that "decision trees in clinical practice may be a suitable alternative to the traditional statistical methods, since it may allow one to analyse interactions between various risk factors beyond the previous knowledge" [251]. Their DT model, based on 194 patients with 9 known clinical indicators, predicted 5-year graft survival with a test accuracy of 74%-88%.

Krikov et al. in their large-scale, multi-centre study [252] analysed 92,844 patient records from the US Renal Data System. Their DT models for long-term kidney graft survival were based on 31 predictors and achieved AUC of 0.63, 0.64, 0.71, 0.82, and 0.90 for the 1, 3, 4, 5, and 10-year predictions, respectively. The trend – the further into the future the forecast scope is, the better its accuracy – appears unintuitive to those working with real-world forecasts. The phenomenon can be explained in part by the way the model accuracies were *measured,* and how this was influenced by reduced follow-up and class imbalance dynamics over the years as more transplants failed.

Decruyenaere et al. compared the traditional logistic regression method with 8 different ML algorithms for the prediction of *delayed graft function* (DGF) following kidney transplantation [253]. Their models were developed on 497 single-centre (Belgium) patients from deceased donors and used 24 parameters related to donor and recipient characteristics, preservation and operation. The authors found that tree-based models achieved low accuracy: AUC of 0.53 for DT and 0.74 for RF respectively, which again can be attributed to DT sensitivity to the high class imbalance between DGF+ (12.5%) and DGF- samples. Out of 10 classifiers, a linear support vector machine performed best with AUC of 0.84.

The models in the above studies were developed with a *few hundred* to a few *tens of thousands* of samples involving national databases. In all four datasets, the number of observations n per p predictor features had ratio of $n/p > 20$. Datasets of such magnitude are not readily available in *HLA-incompatible* renal transplantation, which is inherently a high-risk, low-volume intervention. The data are further limited for smaller transplant units wishing to analyse their samples without having to wait decades until enough procedures are conducted. Hence outcome prediction in AIT often falls under the *small dataset* condition defined in Chapter 3, as $n/p < 10$.

Machine learning from small datasets results in high variablitity among models of the same design. In the previous chapters it has been shown that identical NNs suffer from large discrepancies in their predictions due to random initial conditions, training order and the split between the training and validation samples. Such discrepansies are common for other ML approaches, including DTs. For example, Lofaro et al. attempted to predict, using DTs, chronic graft nephropathy within 5 years post-transplant from 23 clinical indications based on only 80 samples ($n/p = 3.5$) [254]. The authors reported one DT model with $AUC = 0.847$, 62.5% sensitivity, 7.2% false positive rate, and another

tree with $AUC = 0.824$, 81.3% sensitivity and 25% false positive rate. The volatility among the DT trials was not explicitly disclosed, but the two DTs presented showed significant variation in performance and structure, thus casting doubts on the robustness of the overall results. Successful applications of ML to small single-centre datasets for outcome prediction in AIT are presently not known.

The *primary aim* of this research was to confirm *which* donor-specific immunological indicators and at *what levels* were associated with short- and medium-term patient outcomes in HLA-incompatible transplantation. The *secondary aim* was to develop, from the small dataset of complex *pre-transplant* indicators, a *predictive model* for acute transplant rejection that would support the clinical decision process.

5.3 Data: patient and antibody characteristics

The work presented in this chapter examines multivariate associations in the AIT data collected by the clinical collaborators at the renal transplant unit at University Hospitals Coventry and Warwickshire (UHCW). Included in the study were 80 patients (49 female and 31 male) aged between 18 and 68 years (mean age of 41.8 ± 11.6 years) who received HLA-incompatible renal grafts between June 2003 and October 2012. At the time of transplantation 44% of patients had been living with life-limiting end-stage renal disease (ESRD) for 15 years or longer (mean ESRD duration 11.3 ± 8.2 years).

The patient data collected at the UHCW contained information on the type of transplant (living or deceased), the number of HLA mismatches by class (I and II), including particularly dangerous class II HLA D-related (DR) mismatches, patients' progression on ESRD and other baseline characteristics. Additional *immunological* data, including cytotoxic crossmatching and HLA-specific antibody levels by IgG subclass, were gathered through advanced laboratory analyses [167,168]. Flow cytometry and complement

dependant cytotoxic (CDC) crossmatching were performed prior to transplantation. Pan-IgG HLA class I- and class II-specific antibodies were identified in serum obtained before the immunosuppressive treatment.

The *pre-treatment* pan-IgG DSA levels were measured using fluorescence immunoassay and recorded as Median Fluorescence Intensity (MFI) values. In the immunoassay analyses conducted, the positive reactive MFI threshold was set at 1000. The MFI threshold levels for each HLA-specific IgG subclass were five times greater than the negative control incorporated into the immunoassay: 120.6 (IgG_1), 72.0 (IgG_2), 62.7 (IgG_3) and 17.2 (IgG_4) [167,168]. These thresholds presently lack standardisation and vary from centre to centre.

Combined, the following 14 *baseline* (pre-transplant) parameters were established as potential predictor variables:

- 7 *continuous:* single highest pan-IgG DSA MFI level, patient's age (years), ESRD duration (years), and 4 total IgG subclass MFI levels (IgG_1-IgG_4)
- 4 *categorical*: cytometry crossmatch (1=bead, 2=flow or 3=CDC), total number of HLA mismatches between donor and recipient (0-6), the number of class II HLA- DR mismatches (0-2), and the number of previous transplants (0-2)
- 3 *binary:* gender (male/female), the presence of both HLA Class I and Class II DSA (yes/no), and an indicator for donor type (live/deceased)

The transplantation outcomes of primary interest were:

- acute antibody-mediated rejection (ABMR)
- medium-term graft survival/failure

ABMR was defined as acute graft rejection within the first 30 days of transplantation. ABMR was confirmed by renal biopsy in the ABMR+ group ($n = 46$), except in four cases, where anticoagulation was given urgently and precluded a pre-treatment biopsy [167]. The remaining patients ($n = 34$), who did not experience rejection in the first 30 days, were categorised as the ABMR- group. Owing to the advances in AIT, even in patients who experienced acute rejection the graft loss could be prevented by timely intervention. In the UHCW centre, rejection was treated by a combination of immunosuppressive drugs, plasmapheresis, and/or intravenous IgG injections for immunomodulation. Among the 80 high-risk patients, 15 experienced graft failure, 6 died with a functioning transplant, and 59 were still alive with a functioning transplant at the time of this analysis.

Table 5.1 presents the results of the univariate comparison between the patients' baseline clinical and immunological characteristics in the ABMR+/- and the graft survival/failure groups [167]. The null hypothesis of no difference between the groups was tested at 5% significance level using two-tail Fisher exact test for *categorical* variables and the Wilcoxon rank sum test for medians of *continuous* variable distributions. Significant differences between groups ($p < 0.05$) are highlighted in **bold**.

Table 5.1 Baseline clinical and antibody characteristics of transplant recipients [167]

Variable	Rejection (within first 30 days)			Graft outcome (deaths excluded)		
	ABMR+ (n=46)	ABMR- (n= 34)	*p*	failure (n=15)	survival (n = 59)	*p*
Age, *median (range)*	42.5 (18-68)	43 (22-67)	0.83	34 (22-50)	43 (18-67)	***0.003***
Male gender, *N (%)*	17 (37)	14 (41)	0.82	7 (47)	23 (39)	0.56
Prev. transpl., *N (%)*	16 (35)	15 (44)	0.49	10 (67)	36 (61)	0.77
ESRD, *median (range)*	13 (0-29)	10 (0-31)	0.60	7 (0-21)	13 (0-31)	0.13
Living donor, *N (%)*	45 (98)	30 (88)	0.16	15 (100)	56 (95)	0.58
DR mismatch, *N (%)*	38 (83)	27 (79)	0.78	13 (87)	47 (80)	0.72
Total mismatches, median *(range)*	3(1-5)	3 (0-6)	0.13	3 (2-5)	3 (0-6)	0.70

CDC positive, *N (%)*	12 (26)	7 (21)	0.61	8 (53)	10 (17)	***0.006***
Single highest pan-IgG DSA MFI, *median (range)*	6058 (869 –13345)	3492.5 (221-17660)	***0.03***	8987 (775-13345)	3788 (221-17660)	***0.004***
Total pan-IgG DSA MFI, *median (range)*	7797.5 (869-45612)	5134 (306-37084)	***0.01***	11568 (775-45612)	5793 (468-27187)	***0.02***
IgG_3 presence, *N (%)*	22 (50%)	13 (38.2%)	0.36	8 (53%)	24 (42%)	0.56
IgG_3 MFI, *median (range)*	255.5 (76.5-2793)	256.5 (75-1541)	0.63	521 (82-2793)	204 (75-1541)	0.24
IgG_1 presence, *N (%)*	35 (79.5%)	18 (53%)	***0.01***	13 (86.7%)	37 (65%)	0.12
IgG_1 MFI, *median (range)*	2393 (162-24589)	2340 (175-16538)	0.69	6691 (175-24589)	1121 (162-16538)	***0.03***
IgG_2 presence, *N (%)*	24 (54.5%)	14 (41.2%)	0.26	10 (66.6%)	26 (45.6%)	0.24
IgG_2 MFI, *median (range)*	581.7 (87-9472)	952.8 (75-5073)	0.62	1595 (102-9472)	432 (75-4819)	0.08
IgG_4 presence, *N (%)*	24 (52.2%)	12 (35%)	0.17	12 (80%)	23 (39%)	***0.008***
IgG_4 MFI, *median (range)*	113 (24-6505)	30 (17.5-321)	***0.003***	53 (21-135)	35 (17-6505)	0.39
Class I & II DSA, *N (%)*	24 (52)	10 (29)	0.07	8 (53)	24 (41)	0.40
DGF, *N (%)*	12 (26)	4 (12)	0.16	1 (7)	14 (24)	0.28
Rejection, *N (%)*	N/A			10 (67)	34 (58)	0.57

For the ABMR+ patients versus the ABMR- group, significant differences were observed in IgG_1 presence (p=0.01) and IgG_4 MFI levels (p=0.003) [167]. In addition, the patients in ABMR+ group had elevated levels of single highest pan-IgG DSA MFI (p=0.03) and total pan-IgG DSA MFI (p=0.01). There were no significant differences between the ABMR+/- groups with respect to the presence or levels of the DSAs in the two remaining subclasses, IgG_2 and IgG_3, nor in the CDC-positive crossmatch [167]. For the graft failure versus survival groups, the differences were significant in IgG_1 MFI levels (p=0.03), IgG_4 presence (p=0.008), as well as in the single highest pan-IgG DSA MFI (p=0.004), total pan-IgG DSA MFI (p=0.02), age (p=0.003), and the CDC-positive crossmatch (p=0.006).

The univariate analysis (Table 5.1) confirmed the initial hypothesis (Section 5.1) of the association of pre-treatment IgG subclass presence and levels with ABMR and graft failure in AIT. However, further investigation was required to establish whether or not

the effects of IgG_1 and IgG_4 subclasses remained significant in the presence of confounding factors in multivariate space [167].

5.4 Exploratory data analysis

In order to explore the effect of the HLA IgG subclasses on acute ABMR and medium-term survival in the presence of confounding clinical and immunological data, the following multivariate statistical analyses were conducted by the author:

- Cox PH regression for medium-term graft survival
- Logistic regression for acute ABMR

5.4.1 Cox proportional hazards model for graft survival

For medium-term graft survival modelling, early post-transplant outcomes (DFG and rejection) were included in addition to the 14 baseline characteristics described in Section 5.3. The pre-treatment IgG subclass information was considered both as continuous MFI values and as a binary presence/absence indicator (based on the cut-off values declared in Section 5.3). The outcome was modelled as time (in weeks) until the event (graft failure).

Among the 80 patient samples, 6 were excluded due to death-censoring and 3 due to missing values. 57 out of the remaining 71 samples were censored at the study end date (July 2014). Backwards stepwise model selection [139] was used to eliminate variables that did not improve the association with medium-term survival. The final Cox PH model (Table 5.2) comprised 8 variables: number of previous transplants, CDC crossmatch, DFG, single highest pan-IgG DSA MFI, and the presence/absence of the 4 IgG subclass DSAs, out of which only the *highest pan-IgG DSA MFI* and *IgG_4 subclass DSA presence* were

statistically significant (p<0.05). The hazard ratios (HR) revealed that death censored graft survival was significantly worse in cases with positive IgG_4 DSA (HR = 5.8, p=0.035) and elevated single highest pan-IgG DSA MFI levels (HR = 71, p=0.012), both known pre-treatment.

*Table 5.2 Cox proportional hazards model for death censored graft survival. Highlighted in **bold** are p<0.05.*

Variable	p-value	Hazard ratio	95% CI	
			Lower	Upper
Previous transplant	0.125	0.443	0.157	1.253
CDC crossmatch positive	0.598	1.455	0.362	5.855
Highest pan-IgG DSA (MFI)	**0.012**	**70.999**	**2.578**	**1955.4**
IgG1 (+/-)	0.665	0.641	0.086	4.789
IgG2 (+/-)	0.282	0.342	0.048	2.415
IgG3 (+/-)	0.464	1.694	0.414	6.932
IgG4 (+/-)	**0.035**	**5.826**	**1.129**	**30.1**
DGF	0.165	0.225	0.027	1.853

Separate Cox PH analyses were carried out for IgG DSA values at future time points, including peak (around 14th day) and 30th day post-transplant. These time points are of great clinical importance and were considered in this interdisciplinary collaborative study with the resulting models published in [167]. These models, however, relate to post-event information, and, therefore, are not relevant to this thesis, which focuses on *predictive modelling* as a means of stratifying the risk of a particular clinical outcome while it is still beneficial for the patient.

5.4.2 Logistic regression for acute rejection

A multivariate LR was performed to determine whether the pre-treatment IgG_4 and IgG_1 DSA MFI was independently predictive of acute ABMR. Three cases were excluded from this analysis because of missing baseline data, thus leaving 77 samples, 43 in the ABMR+ group and 34 in the ABMR- group.

Included in the LR model were the 14 baseline characteristics outlined in Section 5.3. Numerical variables were mapped to the interval [0, 1] in order to normalise the varying variable ranges. Backwards stepwise model selection eliminated input variables that reduced the quality of the ABMR model. The following five variables were not found to be statistically significant in the LR model:

1) recipient's age
2) ESRD duration
3) number of class II HLA-DR mismatches
4) number of previous transplants
5) the marker of whether the donor was a live/deceased

One explanation as to why these variables did not add value to the ABMR mode is that: both 1) the ESRD duration and 2) recipient's age carry the information on long-term effects, whilst the outcome in question, i.e. acute ABMR, is a snapshot in time at day 30 post-transplant and was influenced predominantly by short-term IgG dynamics; 3) the number of class II HLA-DR mismatches is counted in the total number of HLA mismatches, which was already included in the model; 4) the number of previous transplants correlated with long-term survival, but not acute ABMR; 5) with only 6 cases of deceased donor kidneys (remaining AIT transplantations were from living donors), it is likely that deceased donor incidence was too rare to be captured in the available 80 sample dataset, yet alone used to predict the ABMR.

The resulting LR model shown in Table 5.3 comprised 8 variables and a constant (intercept). The regression coefficients β were statistically significant ($p < 0.05$) for all variables. In particular, 3 variables yielded odds ratios e^{β} beyond the [0.5, 2] interval, namely: the number of HLA mismatches (e^{β} = 4.2, p<0.0001), single highest pan-IgG DSA MFI (e^{β} =3.3, p<0.0001), and total IgG_4 MFI level (e^{β} =3.0, p<0.0001).

*Table 5.3 Logistic regression model for acute transplant rejection. Highlighted in **bold** are odds ratios outside [0.5, 2] interval.*

Variable x	Coefficient β	p-value	Odds ratio e^{β}	95% CI	
				Lower	Upper
Intercept	-1.847	<0.0001	**0.158**	0.089	0.280
Number of HLA mismatches	1.435	<0.0001	**4.199**	2.756	6.396
Class I & II DSA presence	0.386	<0.0001	1.471	1.260	1.718
CDC crossmatch positive	-0.413	<0.0001	0.662	0.575	0.762
Highest pan-IgG DSA (MFI)	1.185	<0.0001	**3.269**	2.557	4.180
IgG_1 (MFI)	0.547	<0.0001	1.727	1.363	2.190
IgG_2 (MFI)	-0.454	<0.0001	0.635	0.598	0.675
IgG_3 (MFI)	-0.235	<0.0001	0.791	0.772	0.810
IgG_4 (MFI)	1.088	<0.0001	**2.969**	2.203	4.000

Taking into account the scaling of the numerical variables, these odds ratios should be interpreted as follows:

- Between the lowest (1) and highest (6) number of HLA mismatches, the odds of the transplant being rejected are expected to increase by 4.2 times.
- For every 1000 MFI units increase in the highest pan-IgG DSAs, there is 13% increase in the odds of ABMR.
- For every 1000 MFI units increase in IgG_4 levels, the expected increase in the odds of ABMR is 30.5%.

In order to establish the *relative* variable importance for ABMR association in the LR model, the likelihood ratio χ^2 significance was evaluated for each variable using Chi-squared test [143]. The analysis revealed that only two variables resulted in a significant (p<0.05) increase in the goodness of fit of the LR model:

- single *highest* pan-IgG DSA MFI (χ^2= 4.3, p=0.003)
- total *IgG_4* DSA MFI (χ^2= 7.6, p=0.005)

The number of HLA mismatches had a high likelihood ratio (χ^2=11.9), but at p=0.06, it fell short of the significance threshold. This could be due to the fact that the Chi-squared test penalises large degrees of freedom, hence this multilevel (6 possible HLA mismatch values) variable was penalised heavily in the likelihood ratio significance test.

Accounting for confounding baseline indicators, the descriptive Cox PH and LR models confirmed the important associations of the single highest pan-IgG DSA and the total IgG_4 DSA subclass with short- and medium-term transplant outcomes [167,255]. The key immunological risk factors established in these exploratory analyses were further confirmed by the predictive ML models described in the subsequent section.

5.5 Predicting early rejection using tree-based learning

As stated at the end of Section 5.2, the primary purpose of this research was not only to establish the key immunological indicators of transplant rejection and its subsequent loss, but to find the baseline *levels* of DSAs for safe transplantation, i.e. establishing *how much* of DSAs can be tolerated before the donor kidney is rejected. The solution to this task required more granularity than had been achieved with the standard Cox PH and Logistic regression analyses, hence it was approached with machine learning.

The secondary goal of the research stipulated a *predictive model* for acute rejection, which clinicians could use to identify patients at risk of acute ABMR *prior* to AIT interventions and to make timely and informed life-saving decisions. Accurately classifying AIT patients into ABMR+/- groups – based on *heterogeneous* (continuous, categorical nominal and categorical ordinal variable types), *multimodal* (routine collection and dedicated laboratory experiments), *incomplete* (a few samples contained missing values), and *small* data – constituted a non-trivial task that also necessitated the use of ML.

Among various ML classifiers, DTs are particularly well suited for clinical classification tasks, where interpretability is key. DTs are easy to interpret by non-statisticians and are intuitive to follow. They cope with missing values and are able to combine heterogeneous data types into a single model, whilst also performing automatic feature selection [121,124]. When combined in a random forest (RF) ensemble, DTs lose part of their interpretability, but benefit from increased robustness and the classification accuracy of RFs. The exploratory power is partially restored in the RF by leveraging the built-in variable importance estimation (Section 2.3.3).

As has been demonstrated in Section 5.2, tree-based learning has been successful in the general area of kidney transplantation where training samples were abundant. The new challenge was to develop equally successful DT and RF models using only a limited, single-centre dataset. The novel application of DTs for the prediction of acute ABMR in HLA-incompatible kidney transplantation was enabled by the *multiple runs* strategy for small data proposed in Chapter 3.

5.5.1 Decision tree and random forest design

From the 80 available patient cases, 60 were randomly sampled for model training and the remaining 20 samples were reserved for independent tests. As has been shown in Section 5.4, the patient groups were well balanced (46 in ABMR+, 34 in ABMR-), but contained 3 samples with partially missing values. In one of the samples the ESRD duration was lost upon collection; in two other samples the IgG_1, IgG_2, and IgG_3 values were not recorded. The 3 missing samples were included in the DT and RF inputs to ensure that the models were able to make realistic predictions on incomplete data, which are commonly encountered in a clinical AIT setting.

The models were evaluated using confusion matrices and ROC curves (as described in Section 2.5); the correct classification rate C, sensitivity Sn, specificity Sp, and area under the ROC curve AUC were evaluated separately for training and test cohorts. The same test cohort was used for both the final DT and RF models.

Decision tree design was based on the standard CART algorithm implemented using MATLAB™ [120] and the Caret package in R [157]. All 14 baseline and immunological predictors described in Section 5.3 were included as the DT input feature space. Throughout the training process, the dataset was recursively divided according to the Gini's Diversity Index split criterion, as described in Section 2.2.3, until the optimal DT hierarchy of nodes was reached. In order to control leafiness, the following constraints defined in Section 2.2.4 were imposed on the DT: minimum parent size of 10, and minimum leaf size of 1. No separate validation cohort was afforded from the already limited number of available samples. Instead, pruning [120] was applied in order to penalise the complexity of the DT and prevent overfitting, thus ensuring that only the most significant splits were discovered by the model.

Using the multiple runs strategy, the experiments with DT were repeated 600 times and each time a different model subset was sampled from the original samples. It was expected that the performance of those 600 DTs would be highly volatile, reflecting that not all DTs would be able to learn from the limited training data. It was also expected that some DTs would be initialised to the training subsets which was more conducive to generalisable patterns in the ABMR+/- patients, as was the case with the small-data neural network models in the concrete and bone applications [232,233]. The sufficient size of the multiple run was estimated from the initial design exploration, where increasing the size above 600 trees did not result in observable changes in the AUC and C distributions.

Random forest design. A RF comprising 600 constituent trees was developed in order to increase the robustness of the stand-alone DT predictions to the degree required by a practical clinical support system. This number of trees was selected due to the same considerations as in the multiple runs above. Although it was expected that the RF would produce substantially more robust results than the 600 individual DTs, the experiment with RF was repeated 10 times to monitor the variance due to small data. In order to reduce its input dimensionality, the predictive RF system leveraged the findings of the exploratory LR analysis conducted in Section 5.4.2. The 5 variables that lacked significant association with the acute ABMR in the LR model were removed from the RF input feature space, thus reducing it to 9 baseline predictors, all of which were known *prior* to the transplantation.

A constituent tree in RF was different from a DT in the following ways:

- Overfitting was controlled by out-of-bag validation with 90% of the samples, as opposed to DT pruning.
- To compensate for otherwise excessively large trees grown without pruning, the minimum number of samples per leaf node was increased to 3.
- Out of the 9 input features, 6 were sampled at random for each partial-feature tree in the RF.

5.5.2 Decision tree model results

The DT classifier in Figure 5.2 was developed after considering a multiple run of the 600 DTs, each modelled on a different subset of the data by permuting the test and training datasets with each other. The models were analysed in a semi-automatic manner whereby the high-performing ($AUC_{train} \geq 0.8$) DTs were monitored for repeating patterns of variables in the branches. In the absence of a separate validation cohort, the

final model was selected as the highest performing (measured by AUC_{train}, C_{test}, C_{train}, and C_{test}) from the subset of DTs with a repeating pattern. As expected, considerable volatility in performance and structure was observed among the 600 DTs. Out of the 600 DTs, a persistent pattern was observed in 14 high-performing DTs, which used the same 6 variables (in differing order) as the model in Figure 5.2.

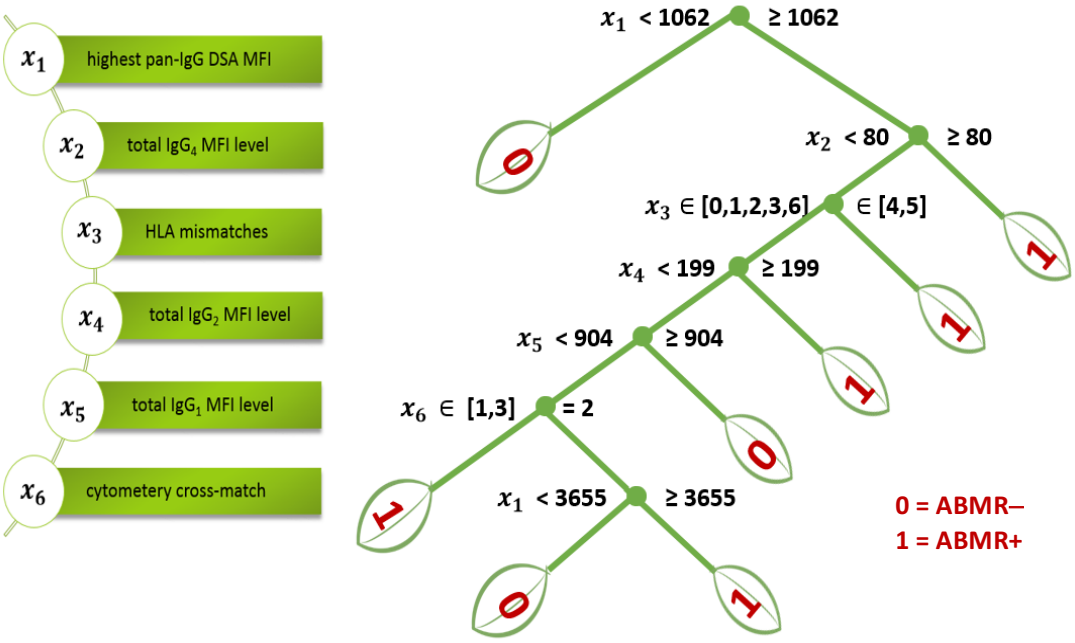


Figure 5.2 DT schematic showing the split hierarchy with 7 branch nodes and 8 leaves based on 6 variables

A further comparison of these 14 DT instances with the remaining 586 DTs in the multiple run was carried out with the Wilcoxon rank sum test for medians of C_{test}. Figure 5.3 shows that despite the overall large variance ($\sigma = 0.013$) across the multiple run, the 14 DTs based on the 6 variables identified in Figure 5.2 had a significantly higher predictive power (p<0.002). This indicated that the training cohorts of these trees, containing high-risk patient groups, were more conducive to learning the associations between the input variables and acute ABMR [232].

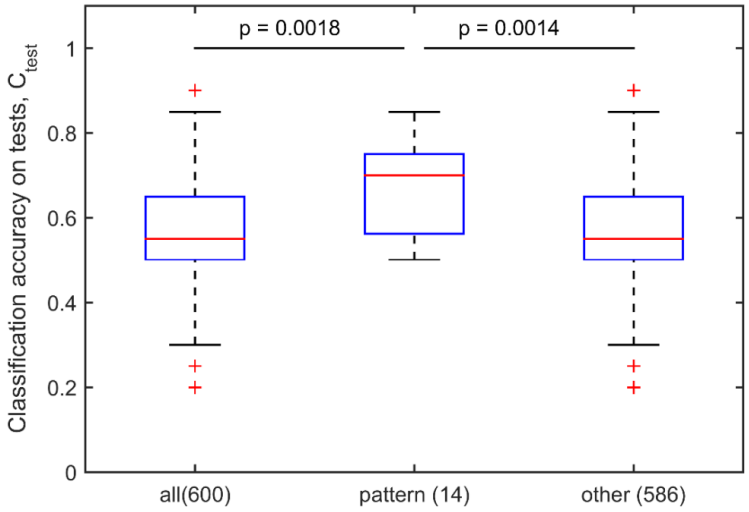


Figure 5.3 Wilcoxon rank sum test for median C_{test} based on 600 DTs and on the subset of DTs with the repeating pattern.

Out of the 14 baseline predictors, 6 variables were identified by the DT as the primary splits for ABMR prediction. These were: the single highest pan-IgG DSA MFI level, total IgG_4 MFI level, number of HLA mismatches, total IgG_2 MFI level, the total IgG_1 MFI level, and cytometry crossmatch. The remaining 8 were not used by the final DT (Figure 5.2) in either the primary or surrogate splits. Thus, the DT model independently confirmed that none of the 5 baseline parameters eliminated by the LR model (Table 5.3) were instrumental to acute ABMR prediction.

Importantly, the node splits in the DT model (Figure 5.2) provided an indication as to what specific *levels* of HLA DSA antibodies were statistically associated with each of the ABMR+/- groups. The DT identified that all patients with the highest pan-IgG levels below MFI 1062 belonged to the ABMR- group (no rejection), while those with the highest pan-IgG level ≥ 1062 and the IgG4 MFI level ≥ 80 had a high (85%) likelihood of early transplant rejection. Similarly, 85% of patients with 4 or 5 HLA mismatches, the highest pan-IgG level ≥ 1062, and IgG4 MFI level < 80 belonged to the ABMR+ group [256,257].

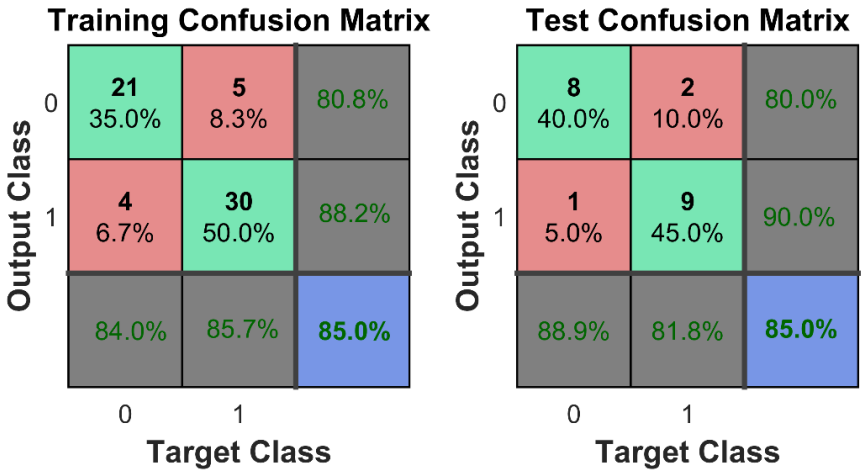


Figure 5.4 Confusion matrices for the training dataset (left) and test samples (right) of the DT model. The cells provide the performance metrics described in Section 2.5.

The DT was able to correctly predict the incidence of ABMR in 85% cases on both training and test datasets (Figure 5.4). When evaluated on the test cohort, the DT identified ABMR+ patients with 81.8% sensitivity and ABMR- cases with 88.90% specificity (Figure 5.4). The classifier ROC curves show $AUC_{train} = 0.849$ on training samples and $AUC_{test} = 0.854$ for DT predictions on test samples (Figure 5.5).

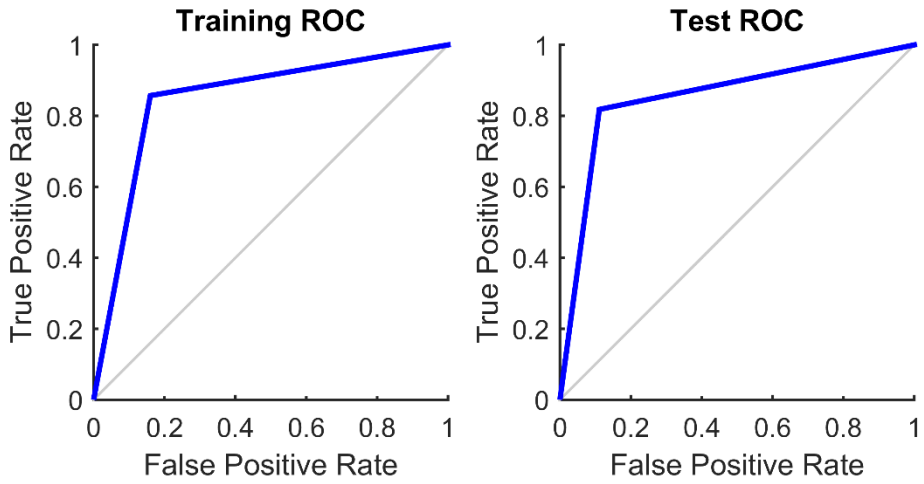


Figure 5.5 ROC curves for DT accuracy on the training dataset (left) and test samples (right).

5.5.3 Random forest model results

The RF of 600 partial-feature trees was built on the same training cohort as the DT presented in Figure 5.2. It achieved C_{train} = 91.7% during the training phase and correctly classified 85% of test cases (Figure 5.6), which was analogous to the DT model performance on the same test cohort. The RF was able to identify ABMR+ patients with a higher sensitivity (Sn = 92.3%) than the DT, but its ABMR- predictions were less specific (Sp = 71.4%). AUC_{test} = 0.819 of this RF was equal to that of the DT (Figure 5.7).

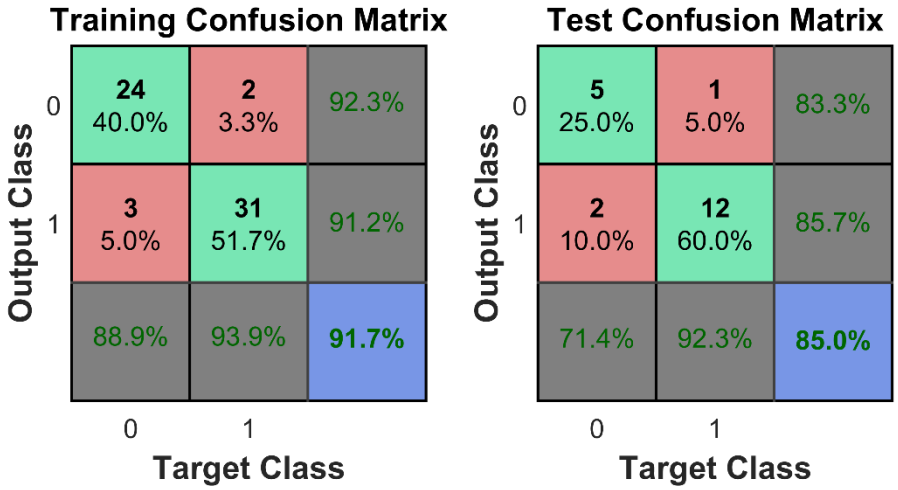


Figure 5.6 Confusion matrices for the training dataset (left) and test samples (right) of the RF model.

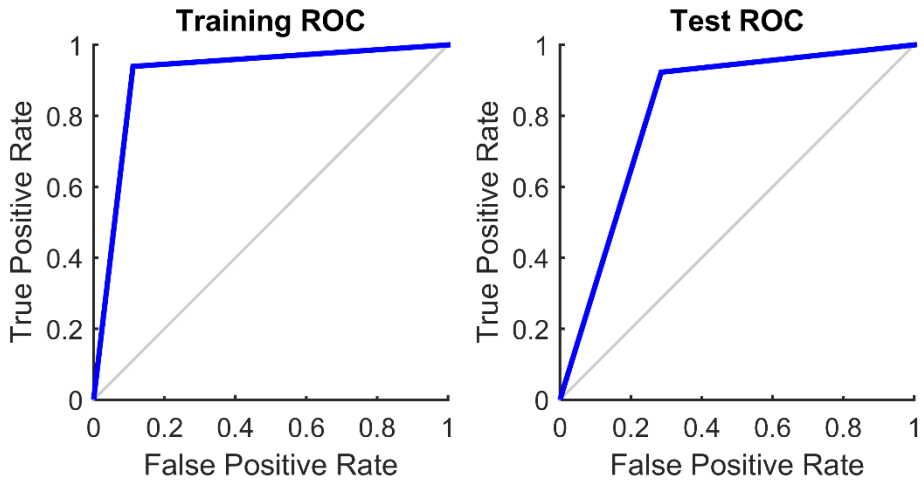


Figure 5.7 ROC curves for RF classification accuracy on training (left) and test (right) samples.

Ten RFs were generated [256] in order to determine whether the consistency of the predictions improved compared to the DT run. The results (Figure 5.8) showed significantly reduced variance ($\sigma = 0.002$), and consistently high performance.

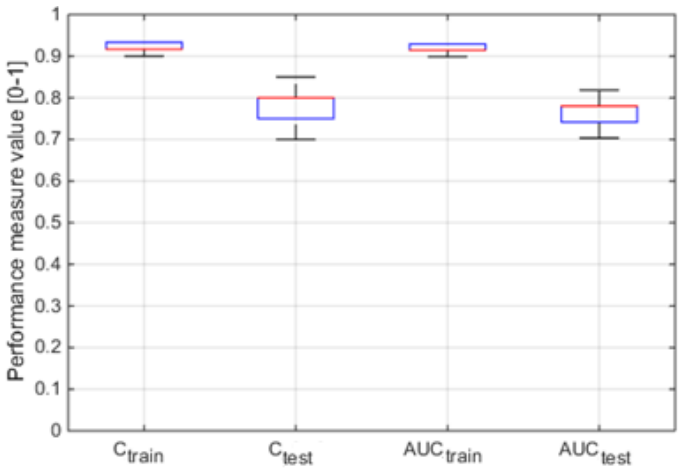


Figure 5.8 Distributions of performance measures C_{train}, C_{test}, AUC_{train}, AUC_{test} for 10 RFs.

The variable importance scores (Figure 5.9) were computed in order to identify the principal factors of ABMR among the 9 input variables used by the RF classifier. As shown in Figure 5.9, the total IgG_4 MFI level was the single most important factor, followed by the highest MFI IgG level, and the number of HLA mismatches. This result independently confirms the finding of the multivariate analyses that IgG_4 was a key contributor to the risk of kidney rejection in the early post-transplant period [255–257].

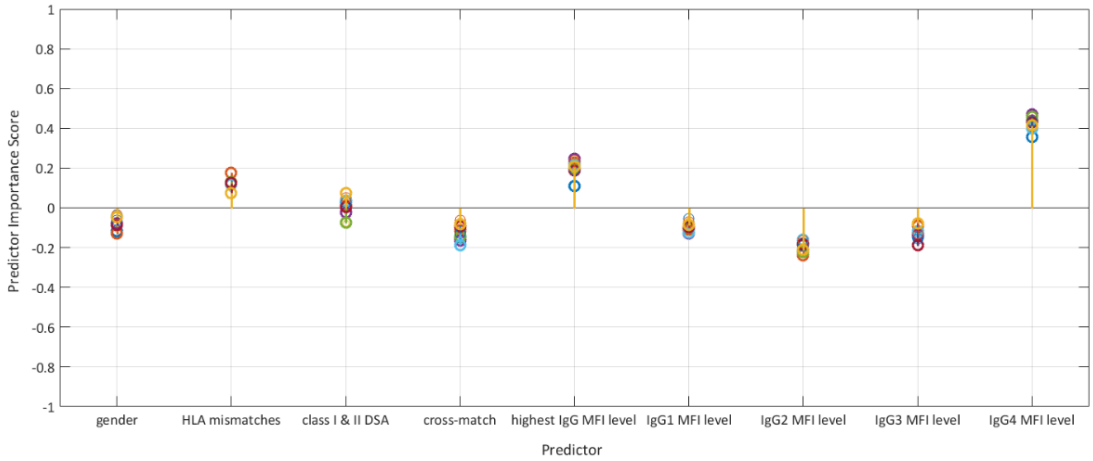


Figure 5.9 Variable importance scores evaluated by a permutation test across 10 RFs

Both the DT and RF models enabled accurate predictions of acute ABMR from the baseline indicators in the UHCW data. However, the two tree-based models offered distinct auxiliary functionality. Despite its volatility to limited training samples, the DT model had an added advantage of *descriptive* modelling: its numerical branches enabled quantification of dangerous HLA DSA levels, whilst its clear graphical representation is easy to follow by non-statisticians. The RF did not offer such ease of interpretation, since it comprised 600 individual partial-feature trees. This drawback in interpretability was compensated by the reduction in the RF performance volatility, making the RF model more suitable as a practical clinical risk stratification system. The quantitative comparison of the two predictive models is provided in Table 5.4.

Table 5.4 Predictive performance of the DT and RF models

Performance measures:	DT		RF	
	training	**test**	**training**	**test**
Correct classification rate, *C* (%)	85.0	85.0	91.7	85.0
Sensitivity, *Sn* (%)	85.7	81.8	93.9	92.3
Specificity, *Sp* (%)	84.0	88.9	88.9	71.4
Positive Predictive Value, *PPV* (%)	88.2	90.0	91.2	85.7
Negative Predictive Value, *NPV* (%)	80.8	80.0	92.3	83.3
Area under the ROC curve, *AUC*	0.849	0.854	0.914	0.819

5.6 Methodological significance and limitations

The accuracy achieved by the DT classifier in Section 5.5 demonstrated that tree-based ML could be effectively applied to predictive modelling in AIT despite the small number of observations and heterogeneous input parameters. Developed with only 60 cases, the DT model for acute ABMR correctly classified 85% of the patients in both the training and test cohorts. Although no similar ML model for acute ABMR existed at the time of

publication [256,257] to make a direct comparison, the proposed DT outperformed in its accuracy (AUC = 0.854) some of the highest-performing large-data ML models in the area of kidney transplantation discussed in Section 5.2 [251–254].

In addition to providing patient-specific ABMR risk predictions, the DT was also a *descriptive* model. Its branch nodes determined the optimal set of 6 pre-treatment indicators associated with acute ABMR, which confirmed and expanded the previous findings of the LR model (Section 5.4.2). The superiority of the DT model was in further granularity: not only had it identified *which* IgG subclasses were highly pertinent to ABMR, but also *what levels* of these IgG DSAs could be safely tolerated [256,257].

The limitation of the DT model was its sensitivity to the subset of training data it received. Without an additional validation cohort it is unknown whether the DT performance on the test cohort was also subset-dependent, or whether it was generalisable to new samples. The model provided in Figure 5.2 was not a unique solution to the classification of ABMR. Instead, it represented one of several DT hierarchies that could explain the association of the samples in the training cohort with acute ABMR. Due to this, no claims on the DT model generalisation for the patient samples outside of the UHCW data could be reasonably made.

The RF ensemble provided an extension to the DT model, with the purpose of improving its robustness as a classification tool. It has been widely accepted that an aggregate vote of several DTs was inherently less noisy and less susceptible to outliers than a single DT output [132–134,138,258]. The RF ensemble in this study was not developed with an intention of improving the already high DT model accuracy, but to factor in possible noise in the training samples the DT model received, and thus increase its generalisation potential. The RF offered a better consistency of results and lowered the volatility of the DT predictions, albeit at the expense of reduced interpretability. It remains for further

study to confirm that the reduction in volatility was due to ensemble learning and not the training data, and whether the RF model would indeed be able to generalise on the patient cohort outside of the UHCW centre.

Small dataset size was not the only limitation of the clinical data explored in this work. Despite being meticulously collected and maintained by largely the same team of nephrologists, the single-centre dataset contained 3 incomplete samples. These samples with partially missing baseline and immunological information could not be integrated into the Cox PH survival and logistic regression analyses without one of the imputation strategies discussed in Chapter 3. On the contrary, the DT and RF models were well-equipped to handle partially missing data and managed to classify correctly all 3 cases with incomplete data. This methodological superiority of the tree-based ML models further adds to their descriptive and predictive significance for classification from limited clinical data.

5.7 Clinical impact

At the time of publication [167,232,233], this research was the first in the UK to use machine learning for the prediction of acute ABMR from HLA donor-specific IgG subclass data in antibody-incompatible renal transplantation. It was also the first demonstration of the potential prognostic value of the HLA DSA IgG_4 in AIT [167,255,257].

The independent association of IgG_4 DSAs with the graft outcome was first confirmed by the Cox PH model (Table 5.2). The multivariate model revealed a strong association of pre-treatment IgG_4 DSA *presence* with medium-term graft loss. Accounting for multiple confounding factors, this association was independent of IgG_1 and pan-IgG DSA levels, revealing that IgG_4 DSAs, even in isolation, could be highly pathogenic to the graft. It is also possible that, being last in the IgG class-switching sequence, IgG_4 represented an

already mature immune response to the donor organ by the recipient's immune system. The importance of IgG_4 was also detected in acute graft rejection by the exploratory logistic model (Table 5.3). The LR model confirmed a significant association of pre-treatment IgG_4 DSA *levels* with acute ABMR. A similar association was independently confirmed by the relative variable importance scores in the subsequent RF model. Combined, the analyses strongly supported the discovery of IgG_4 DSAs as the key prognostic indicator for short- and medium-term AIT outcomes.

Conventional statistical models were unable to determine how much of IgG_4 DSAs could be safely tolerated before the transplant was rejected by the recipient's humoral system. It was the DT model that revealed the dangerous *levels* of antibodies associated with ABMR. The harmful levels of IgG_4 DSA and the single highest pan-IgG DSA were identified to be at 80 MFI and 1062 MFI, respectively (Figure 5.2). Whilst the threshold of around 1000 MFI for the single highest pan-IgG had been intuitively used by transplant experts [75,239,248,259], the threshold for IgG_4 discovered by the DT model on the UHCW data presented an entirely novel insight.

By integrating known and novel associations, the tree-based ML classifiers developed in this work enabled accurate, patient-specific outcome predictions for acute ABMR. They provided the means for the early stratification of ABMR risk from pre-treatment clinical and immunological indicators, leaving clinicians with more time to make essential adjustments to treatment. The granularity, with which the DT model determined *which* IgG subclasses were particularly dangerous, and to *what degree*, added invaluable statistical evidence to support the expert clinician's decision making. These outcomes are summarised by the workflow schematic in Figure 5.10.

By informing clinical decisions, tree-based ML has the potential to transform personalised care in AIT, preventing life-threatening graft loss and over-treatment by

costly and harmful immunonosuppresive drugs. Before the DT and RF prototypes developed in this work on the UHCW data can be used as a practical decision support tool, they require extensive validation with external datasets. Through dissiminations at multiple international conferences and leading AIT fora, requests to collaborate were discussed with multiple groups, including the Paris centre [260]. A grant application to obtain additional HLA-incompatible and blood-group incompatible transplant data was submitted to the UK Transplant Registry [261] and access has recently been granted. Thus, the research underpinning this thesis forms the foundation for an extensive multi-centre collaboration with the potential to transform the field of antibody-incompatible renal transplantation.

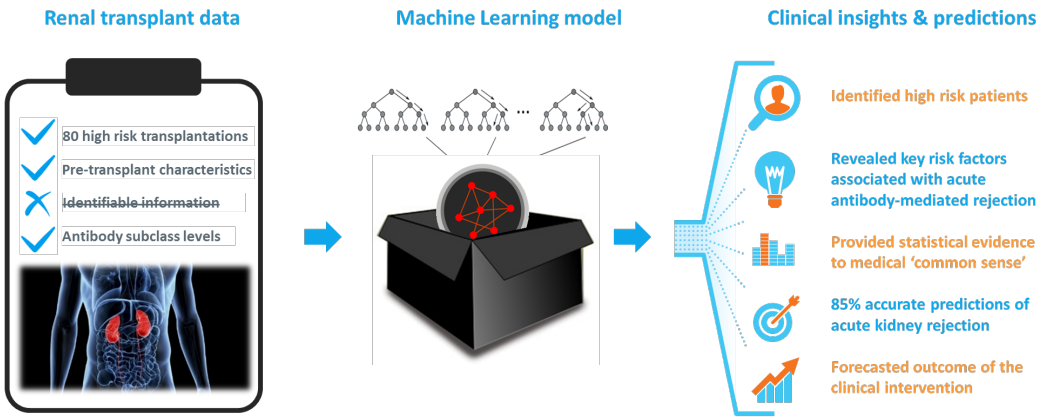


Figure 5.10 From raw data to clinical insight: summary of the workflow

5.8 Chapter conclusions

The key finding of this chapter are as follows:

(1) Single-centre renal transplant data were explored for novel multivariate associations and the potential for data-driven predictive modelling of early transplantation outcomes.

(2) A multivariate Cox PH model established the independent association of the single highest pan-IgG DSA MFI levels (HR = 71, p=0.012), and, specifically, IgG4 DSAs presence (HR = 5.8, p=0.035) with medium-term graft loss.

(3) An exploratory LR model confirmed that the single highest pan-IgG DSA MFI level (e^{β}= 3.3, p<0.0001) and total IgG4 DSA MFI level (e^{β}= 3.0, p<0.0001) were also associated with early transplant rejection.

(4) A predictive DT model, developed on 60 patient samples using the method of multiple runs, independently confirmed the confounding factors used in LR and predicted early ABMR with 85% accuracy.

(5) By providing a quantification of dangerous DSA levels, the DT identified that all patients with the highest pan-IgG levels below MFI 1062 belonged to the ABMR- group, while those with the highest pan-IgG level ≥ 1062 and the IgG4 MFI level ≥ 80 had an 85% chance of early ABMR+.

(6) Within the limitations of the test cohort, the predictive RF ensemble model improved DT robustness and maintained 85% predictive accuracy. The relative variable importance scores of the RF ensemble further confirmed the LR and DT findings of the key immunological and clinical factors for acute ABMR in HLA sensitised patients.

Chapter 6

Diabetes type 2 risk stratification from routinely collected NHS data

The research presented in this chapter stems from a 3-year-long collaboration with the Nuffield Department of Primary Care Health Sciences, University of Oxford. The overall aim of the project was to explore the possibilities of improving the existing type 2 diabetes risk stratification system used in NHS primary care, through the adoption of ML and inclusion of blood glucose information. The author's contribution to this collaboration was the development, implementation, and validation of a novel ML prototype that predicted the 10-year risk of acquiring type 2 diabetes in the UK population based on routinely collected primary care data. The study protocol specified for the model to be based on artificial neural networks, although alternative models using logistic regression and survival decision trees were also explored and are presented in this chapter.

6.1 Diabetes in the UK and globally

Diabetes mellitus (DM) is a chronic hormone deficiency condition that significantly impacts on the lives of an estimated 422 million people globally [262]. The key feature of DM is the relative or absolute absence of insulin – the hormone involved in controlling

and critically lowering levels of glucose in the blood. Elevated levels of blood glucose (hyperglycaemia) produce serious short-term as well as long-term complications that have significant impacts on the quality of life and health of diabetic patients and lead to increased mortality [263]. In severe cases, patients can suffer from hyperglycaemic hyperosmolar state, and in instances of absolute insulin deficiency, lead to ketoacidosis, loss of consciousness and coma. In cases of long-term hyperglycaemia, the microvasculature of a patient's kidney, eye, nerve, and larger arteries are affected, leading to blindness, neuropathy, and end-stage renal failure [263]. Hypoglycaemia (low blood glucose) caused by improper glycaemic management in diabetic patients can also lead to mortality [263]. The World Health Organisation estimates that DM and its complications caused 1.5 million deaths in 2012 (2016). In England alone, DM is responsible for over 48,000 hospital admissions and 5,500 deaths annually [264].

As alarming as its complications, is the accelerating rate at which DM continues to strike modern society. The worldwide prevalence of diabetes type 2 has doubled since 1980 and this trend is expected to continue with a forecasted 592 million diabetic people by 2035 [262,265]. Weber and Narayan call it the “epidemic of diabetes” (2008). In the UK, the number of adults living with DM is 3.9 million, corresponding to a prevalence rate of 6.2% [267,268]. It is estimated that there are also around 850,000 people in England who have diabetes but have not been diagnosed [267]. Those undiagnosed patients may have experienced diabetic complications, such as a heart attack or renal failure without warning symptoms.

The rapid growth in DM incidence is a serious public health priority [262,264]. Fortunately, the past few decades have also brought about two paradigms: (1) widespread computerisation of medical systems in industrialised countries that have resulted in the collection of vast (and often convoluted) digital repositories of patient

data, and (2) the emergence of hardware and machine learning algorithms capable of dealing with these data. Combined, these technological advances offer an unprecedented opportunity to better understand, monitor and manage DM. By analysing routinely collected primary care data using NNs and DTs, this work contributes to the nascent niche for machine learning application in the early prediction of DM in the UK population.

6.1.1 Disease pathology, diagnosis and treatment

Despite the rapid growth in incidence of DM in the past century, the disease is not a phenomenon unique to modern society. The term *diabetes* can be traced back to the 2nd century AD Greece, and the distinction between the different types of DM is found as early as the 5th and 6th centuries AD in the work of multiple Hindu physicians [263]. Yet our understanding of the complex metabolic and biochemical processes in DM, as well as the policy around diagnosing and managing DM, is still evolving.

DM is diagnosed on the basis of chronic hyperglycaemia determined by a blood glucose (BG) test. A fasting glucose level ≥7 mmol/L in plasma or ≥6.1 mmol/L in a capillary blood sample define DM. Where fasting BG is not available, 2-hour BG level of ≥11.1 mmol/L (either in plasma or capillary) are used as a sufficient diagnostic criteria for DM [263]. In 2011 the World Health Organisation recommended a supplementary diagnostic measure based on glycated haemoglobin (HbA1c) above 48 mmol/mol or 6.5% [267].

DM is categorised into two main types: type 1, caused by autoimmune responses within the pancreas (absolute deficiency), and type 2, associated with insulin resistance and impaired secretion. There are a range of other forms of DM, such as neonatal DM, mature onset diabetes of the young (type MODY), and Alström and Wolfram syndromes. However, over 90% of all incidences of DM correspond to type 2 [264]. This work on routinely collected data from the general UK population focuses solely on type 2 DM.

Type 2 DM is caused by a combination of genetic and environmental factors, such as low physical activity, imbalanced diet, chronic stress and polygenic inheritance patterns. It is a life-long condition that can rarely be reversed, although the risk of complications and the severity of the disease can be considerably reduced by life style interventions, such as weight loss, increased physical activity, and cessation of smoking [265]. These life style changes have to be adopted early and carried out persistently to be effective, thus timely detection of DM or pre-DM conditions is highly advantageous [264,266,269].

Unlike type 1 DM, type 2 diabetic patients are not necessarily dependent on external insulin. However, type 2 DM is a progressive condition that often deteriorates, to the point where exercise and healthy diet alone are not sufficient to control BG levels. The patient is then prescribed insulin and a number of medications to stimulate and protect insulin-producing cells in the pancreas or to inhibit the absorption of starch in the intestine and the absorption of glucose by the kidneys and the blood [268]. The ability to maintain BG outside the dangerous hyper- or- hypoglycaemia thresholds is critical to reducing the risk of serious macro- and micro-vascular complications and death [263,269,270]. The longer the patient's body is exposed to uncontrolled hyperglycaemia, the higher the risk of the irreversible damage of insulin-producing cells [263]. Early detection of DM is therefore key to preventing severe morbidity.

6.1.2 Managing type 2 diabetes risk in primary care

With 49.1 million items prescribed each year for DM in England alone, the burden of monitoring and managing type 2 DM falls on primary care [264]. The risk of type 2 DM can be identified through correlated (but not necessarily causal) factors observable in primary care, including:

- high BMI and obesity

- first-degree relative with DM
- smoking
- hypertension
- conditions requiring the prescription of corticosteroids
- gestational diabetes

The net effect of these risk factors is not straight forward. Obesity in itself is a summary measure of multiple health conditions that reflect both lifestyle and genetic factors [265]. Cardiovascular conditions such as hypertension and DM are mutually-cofounded: patients with hypertension have increased insulin resistance, while 75% of DM patients also have hypertension [271]. Some risks are associated with patient demographic. For instance, among the UK population type 2 DM is found to be more common in people of South Asian, African, Afro-Caribbean and Chinese family origins and in people from regions associated with a high Townsend index of multiple deprivation [264]. General practitioners are encouraged to screen for DM risk factors, and to refer patients for a blood biochemistry test when a combination of multiple DM risk factors is observed [264,267].

Population screening for high-risk groups is set in place across UK primary health care practices [264,267]. The official guideline of the National Institute for Health and Care Excellence [267] recommends the use of computer-based risk assessment tools based on routinely collected data, including the QDiabetes® risk calculator [272], the Cambridge dataset risk score [273], the questionnaire-based Finish Diabetes Risk Score [274], and the Leicester practice score [275]. These four risk assessment tools predict the 10-year risk of being diagnosed with type 2 DM and use 5 common variables (age, gender, BMI, family history of diabetes and hypertension) and 2-5 additional variables, such as waist

circumference, self-reported fruit and vegetable intake and gestational diabetes, all of which can be measured in primary care without expensive laboratory tests.

It is unrealistic that a handful of predictor variables can capture the entirety of the possible DM risks and variance in population, but it is important to establish the small number of variables that account for most of the variance. In a systematic review of 14 type 2 DM risk-prediction models, Noble et al. advised against including more than 10 components to the risk model in order to sustain its usability [276,277].

The Finnish Diabetes Risk Score and Leicester risk scores use questionnaires that sort patients into appropriate categories of risk of developing type 2 DM in a 10-year period. The Cambridge risk score and QDiabetes® use data already available in primary care systems, but the algorithms for computing the risk differ: QDiabetes® uses the Cox proportional hazards model to compute percentage risk of developing type 2 DM, while the Cambridge risk score utilises logistic regression to express the likelihood of having undiagnosed diabetes.

Importantly, these four validated and routinely used risk assessment algorithms were recently found to produce dissimilar risk scores [278]. For an individual, this carries an implication that their predicted risk is dependent on which risk-prediction tool is used and could be altered if a different assessment is adopted. The National Audit Office exposed that a high-quality randomised controlled trial was yet to confirm that the existing manual screening is beneficial [277].

The *aim* of this work is not to develop yet another type 2 DM risk score, but instead to investigate the utility of various algorithms, both from classical statistics and machine learning, in the context of routinely collected data.

Of the four risk assessment tools recommended by NICE, QDiabetes® has by far the largest model derivation and validation base. It was built on 2.5 million medical records, amounting to 16,436,135 person years of observation, during which 78,081 new incidences of type 2 DM occurred [272]. The model considers 11 predictor variables: patient's age, gender, gestational DM, BMI, smoking status, self-assigned ethnicity, family history of DM, Townsend multiple deprivation score and whether or not the patient was treated for hypertension, had cardiovascular disease, or was prescribed corticosteroid drugs.

Despite their uncontested diagnostic value, BG measurements are not included in the computation model of QDiabetes®. The reason why QDiabetes® disregards this essential biochemical factor is because it was designed with a vision to be used both in a primary care environment and by patients at home, where a blood test for measuring BG levels may not be available. This assumption is now obsolete: in the 24 years since the collection of the first patient record in the QDiabetes® study, point-of-care testing, including that of BG in primary care, have become routine [279,280]. For a model designed to predict DM risk at the point of care, it would be reasonable to use all available data, including BG biochemistry, even if this will render the model less useful outside the clinical setting. The advances in point-of-care testing and BG monitoring, coupled with the recent trends in personal health devices [280–283] further stipulate the inclusion of BG measurements in the DM risk stratification systems of the future. The models developed in this chapter consider BG information, where available.

6.2 The data

6.2.1 Overview

The data acquisition was funded by the National Institute for Health Research in accordance with the study protocol [90]. The data were obtained through the Clinical Practice Research Datalink (CPRD) and stored on the University of Oxford servers. The data contained information on the incidence of type 2 DM diagnosis and associated baseline indicators spanning a 20-year period from 1/1/1993 to 31/10/2013. 100,000 anonymised EMR were requested from primary care practices; these were randomly distributed within CPRD in order to best represent the wide UK demographic. Patients who had already been diagnosed with DM (either type 1 or type 2) were excluded from the study. Consistent with QDiabetes®, only patients aged 25-79 at the date of entering the study were considered. After applying the exclusion and inclusion criteria, the final study dataset comprised of 79,959 records, totalling 476,333 person-years.

In addition to the variables considered in QDiabetes® [272], the dataset also included biochemical data, although the actual BG levels (fasting or otherwise) were largely incomplete or obsolete (collected more than 5 years prior to the index date). The incidence of gestational diabetes among women in the study was less than 0.01% providing too few (77) events to be reliably included in the model. Hence, the following 11 variables were used as model input:

- continuous: patient's age (years) and Townsend score (dimensionless), most recent at the index date BMI (kg/m^2) and BG (mmol/L) measurements
- *binary*: presence (1) or absence (0) of diabetic family history, incidents of cardiovascular disease (CVD), treatment for hypertension, prescription of corticosteroids (steroid), and smoking history (smoker)

- *nominal*: gender (1=Male, 2=Female) and ethnicity (1=White, 2=Asian, 3=Black, 4=Mixed, 5=Other)

The outcome variable was a binary result of type 2 DM diagnosis at the end of the study (0= non-diabetic, 1=diabetic or missing). The duration (days) from the index date to the study end date or to the type 2 DM diagnosis date was also considered in survival analysis. It is important to note that the *diagnosis* of type 2 DM is merely an indirect measure of whether the person did or did not have type 2 DM. The cases of undiagnosed DM abound as illustrated in Section 6.2.2.4.

Table 6.1 CPRD data: descriptive statistics across the derivation and validation cohorts

Variables *Statistic*	Derivation cohort	Validation cohort
Patient *N*	53306	26653
Person years *Mean (std)*	6.0 (3.7)	5.9 (3.7)
Diagnosed type 2 DM *N*	1585	828
Gender female *N (%)*	26608 (49.9)	13260 (49.8)
Age (years) *Mean (std)*	44.5 (14.7)	44.4 (14.7)
BMI recorded *N (%)*	30722 (57.6)	15236 (57.2)
BMI (kg/m2) *Mean (std)*	26.2 (5.1)	26.2 (5.1)
Any blood glucose recorded *N (%)*	13879 (26.0)	7021 (26.3)
Fasting BG (mmol/L) *Mean (std)*	5.1 (0.8)	5.1 (0.8)
Random BG (mmol/L) *Mean (std)*	5.1 (1.0)	5.1 (1.0)
Townsend score *Mean (std)*	-0.5 (2.9)	-0.5 (2.9)
Ethnicity		
White *N (%)*	50755 (95.2)	25416 (95.4)
Asian *N (%)*	1308 (2.5)	626 (2.3)
Black *N (%)*	564 (1.1)	262 (1.0)
Mixed *N (%)*	265 (0.5)	147 (0.6)
Other *N (%)*	414 (0.8)	202 (0.8)
Smoker *N (%)*	12383 (23.2)	6292 (23.6)
Family history of DM *N (%)*	3331 (6.2)	1678 (6.3)
History of CVD *N (%)*	2157 (4 .0)	1100 (4.1)
Treated for hypertension *N (%)*	5288 (9.9)	2604 (9.8)
Prescribed steroids *N (%)*	1346 (2.5)	699 (2.6)

The data were divided randomly into derivation and validation cohorts: 1/3 of the records were held for the purposes of model validation (also referred to as 'test' cohort),

and the remaining records were made available for model development. Descriptive statistics across the derivation and validation cohorts are presented in Table 6.1.

This CPRD data were gathered by general practitioners and practice nurses during routine patient visits and reflect the broader challenges of routinely collected data. The sources of this complexity are discussed in Section 6.2.2 with illustrative examples from the 79,959-sample dataset used in this study.

6.2.2 The 4 "C"s of routinely collected data

6.2.2.1 Complexity

Routinely collected data carry broad, often overlapping and at times contradictory information about a patient's health and is inherently complex.

The variety of possible underlying physiological interactions between causal, correlated and cofounded factors in life-long conditions such as type 2 DM are not fully established. Type 2 DM presents an intricate interplay of genetic predisposition and metabolic processes, where unhealthy diet, smoking, obesity, and physical inactivity combine with previous gestational DM, ethnicity, and older age. The direct and indirect indicators recorded in the electronic medical system at the point of care may not capture all of this complexity. This is evident from retrospective analysis of large patient databases, where a number of patients may match across all variables of interest, yet their 10-year DM outcome may differ.

For example, among the records in this study, there were two 31-year old white male, non-smoking patients with an almost identical healthy body mass (BMI of 19.2 and 19.7), from equally prosperous demographic areas (Townsend decile score of –2.17) and identically absent histories of hypertension, cardiovascular disease, family DM, previous

BG measurement or treatment with steroids at the time of joining the study. Despite their profiles matching across all 10 of the parameters considered, the two men experience opposite outcomes: one goes on to develop DM and is diagnosed with type 2 DM after 3.5 years, while the other leaves the study after 10 years without DM. Such apparent contradictions are abundant in the dataset, indicating from the start of the study that the variables available for analysis do not contain all the necessary information required for modelling the disease.

6.2.2.2 Completeness (or the lack of)

Routinely collected patient data could be described as *sparse,* with values missing across a range of variables. The level of completeness depends on the mechanism through which the data were collected. For instance, in the CPRD dataset [90] the patient's date of birth and gender were known in all instances, since the medical record would not have been instantiated without the two variables. For variables that represent conditions or comorbidities that require diagnosis or clinical intervention, such as CVD, treatment for hypertension, or prescription for steroids, missing values imply their absence and could be reliably substituted by zero.

More uncertainty is present around variables that rely on patient disclosure, such as ethnicity, smoker status or family history of diabetes. Among 79,959 records, 13% did not have smoking status on record. Family history of diabetes was missing in all but positive cases (9%). Ethnicity was not recorded in 70% of patients. Customarily, these missing values would be left as missing, imputed statistically or the entire patient record would be omitted from the study [17,284,285]. Instead, in this collaborative study, domain knowledge of practicing healthcare professionals and clinical statisticians [90,286–288] was enlisted to deduce the missing values based on the mechanisms by which the samples were recorded. It was decided to treat missing family history of DM

as absent (0), unrecorded ethnicity as "White" (1), and to give benefit of the doubt to the unknown smokers (0 if missing). These subjective assumptions generate noise in the data, which adds to the challenge of modelling with limited and uncertain information.

The final category of missing input variables are continuous variables such as BMI, and blood plasma glucose level: fasting (FBG) or random (BG). Any BMI, FBG and BG measurements available up to 5 years prior to the index date were included in the analysis. Despite this generous threshold, the measurements were grossly missing: BMI indications were absent in 43%, BG - in 79%, and FBG – in 93% of the patient records (Figure 6.1).

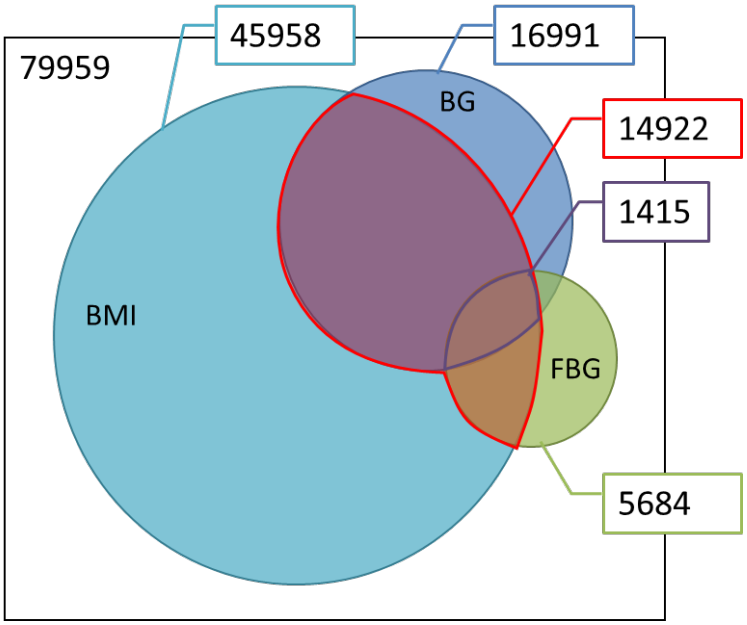


Figure 6.1 Venn diagram representing the number of recorded BMI (aqua), BG (blue), FBG (green) and the union of BG and FBG (red contour) as a proportion of the entire available dataset (white rectangle). The diagram is annotated with the actual numbers of the records corresponding to each subset. The areas of the figures are drawn to scale.

Evidently, the FBG was too scarce to be reliably incorporated into the model on its own. Furthermore, for a healthy patient a *fasting* blood glucose test is expected to yield a more conservative mmol/L value than a measurement taken at an arbitrary point in the day,

yet it can be noted from Table 6.1 that the mean and standard deviation of the BG and FBG are surprisingly similar. A closer analysis revealed that indeed the distributions of the BG and FBG (Figure 6.2) are identical (p-value > 0.05, Wilcoxon rank sum test)

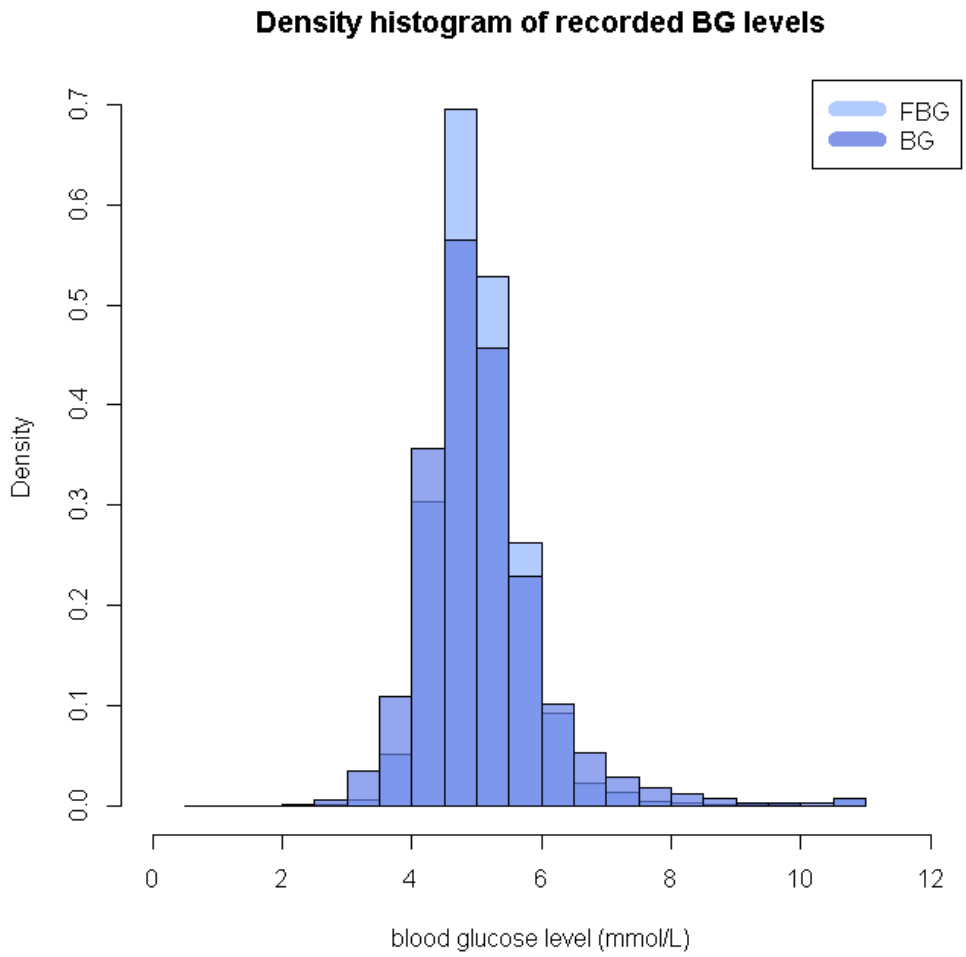


Figure 6.2 Distributions of FBG and BG values

Hence BG and FBG were aggregated into a single predictor, such that:

$$BG_{new} = \begin{cases} FBG, & if\ FBG\ is\ known \\ BG, & if\ FBG\ is\ missing, BG\ is\ known \\ undefined, & if\ neither BG\ nor\ FBG\ are\ known \end{cases},$$

yielding a new variable with a mean of 5.1 mmol/L with a standard deviation of 0.9 mmol/L. This new BG was known for only 26% of the patients; a separate variable, BG_{pres}, was set to 0 for the remaining 74% of patients for whom no BG nor FBG levels

were known. When known, BG levels were significantly different (p-value < 0.0001, Figure 6.3) in patients diagnosed with type 2 DM by the end of the 10 years (diabetic outcome group, median BG 5.8 mmol/L) and those who would have exited the study without a diabetes diagnosis (non-diabetic outcome group, median BG 5.0 mmol/L).

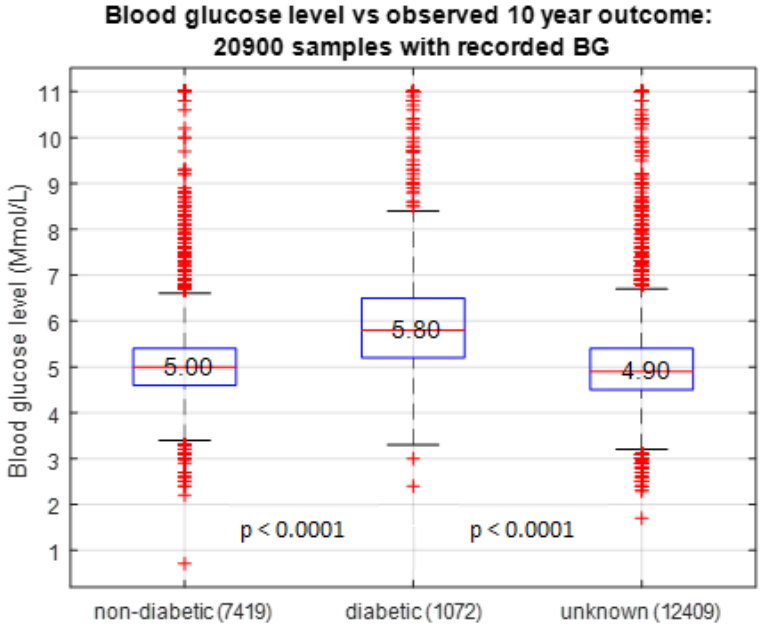


Figure 6.3 Wilcoxon rank sum test for medians for BG values in DM, non-DM and unknown outcome groups.

6.2.2.3 Censoring

Another type of missing data is where the outcome variable is unknown. In primary care, this may happen when a patient leaves the practice before the end of the longitudinal study. The loss of follow-up is arguably the most defining characteristic of routinely collected data. Over 63% of the patients studied in this work left the study before the end of 10 years. The CPRD dataset contained the date of when the patient left the practice and the variable "reason for transferring out", which, among others, included death. It is *unknown* whether those transferred-out patients (apart from those who died) would have been diagnosed with type 2 DM or remained disease-free by the end of the 10 years.

Loss of follow-up and, therefore, the ability to ascertain what outcome would have been developed by a patient if they remained in the study, results in *right-censored* records. Censoring further reduces the number of records for conventional modelling with supervised machine-learning algorithms, which inherently rely on known outcome labels for training. Let us recount the proportion of available samples with BMI and BG present in the Venn diagram in Figure 6.1, taking into account censoring. The resulting Figure 6.4 demonstrates the "big picture" of the combined effect of censoring and missing BMI and BG measurements on limiting the samples available for supervised modelling.

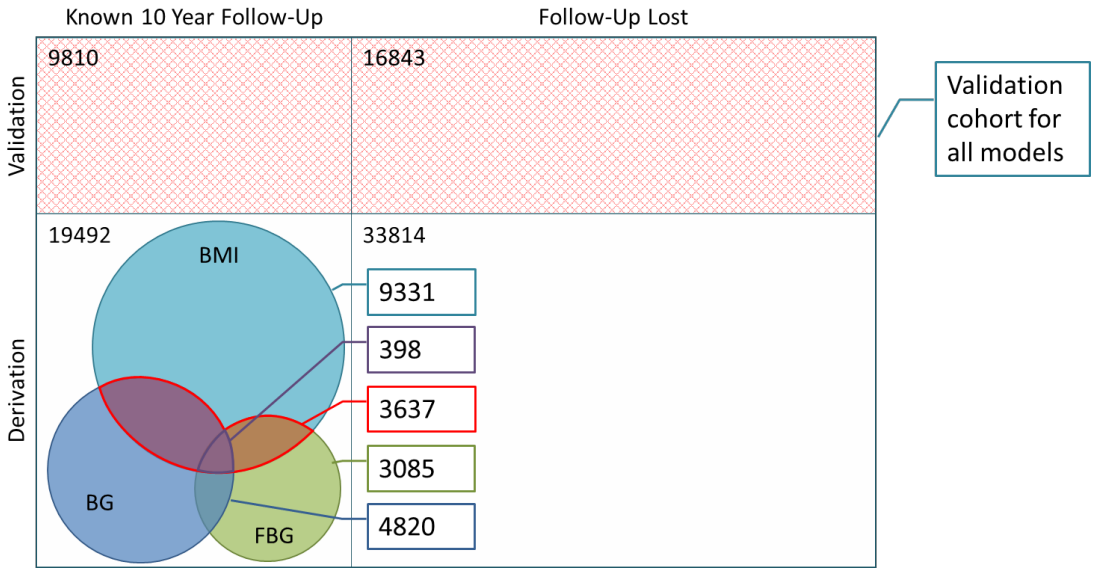


Figure 6.4 Area diagram representing the number of samples with known 10-year follow-up (left) and unknown outcome (right), separately in validation (top) and derivation (bottom) cohorts. The areas of the rectangles are drawn in proportion to their populations.

6.2.2.4 Consistency

Despite numerous attempts toward standardisation, the distributed nature of general practices makes primary care data prone to institutional bias: from one clinic to another (inter-institutional bias), from one nurse to another (intra-institutional bias), or from one day to the next. This determines *if* and *when* certain baseline characteristics are recorded [285]. Moreover, *how* those indicators are interpreted also depend on when,

where and by whom the patient was seen [289]. Whilst it is not the primary goal of this research to quantify the various inter- and intra-centre biases, it is important to acknowledge that inevitable inconsistencies influence why the likelihood of BG and BMI values being recorded differs among DM and non-DM groups, and why some patients with BG levels meeting diagnostic criteria remain undiagnosed for several years.

Firstly, the existence of bias in a variable being recorded for patients of different outcome groups directly affects the utility of statistical techniques such as multiple imputation on that variable (Section 3.1.3). Table 6.2 illustrates that patients who went onto develop type 2 DM were found 1.5 times more likely to have their BG levels measured. Existence of a BG record was biased towards patients in the DM-outcome group versus non-DM and unknown outcome groups. Whilst the mechanism by which a patient might be referred for a BG test was unobservable from existing data, the bias indicated that BG variable violated the Missing at Random (MAR) assumption required for multiple imputation. On the other hand, the existence of a BMI record was as likely in patients who would develop type 2 DM as in patients who transfer out without a diagnosis. The absence of a significant bias, coupled with a larger proportion of known values (57%) provided sufficient ground to impute BMI.

Table 6.2 Frequency of BG and BMI being recorded for the whole cohort and separately for DM, non-DM and unknown outcome groups.

Cohorts:	**all**	**DM**	**non-DM**	**unknown**
Is BG measured uniformly among DM, non-DM and unknown outcome groups?				
# samples	79959	2413	26889	50657
# recorded BG	20900	1072	7419	12409
%	26%	44%	28%	24%
Is BMI measured uniformly among DM, non-DM and unknown outcome groups?				
# samples	79959	2413	26889	50657
# recorded BMI	45958	1463	12576	31919
%	57%	61%	47%	63%

Secondly, patients with BG levels meeting diagnostic criteria may remain undiagnosed due to inconsistencies in the interpretation of already existing records and/or the lack of monitoring of patients at risk. Recent studies demonstrated that cases of *undiagnosed type 2 DM* were more common than previously acknowledged [286,287]. Among the patients analysed in this work, 20900 individuals had either their BG or FBG levels recorded at least once within 5 years before entering the study. The records indicate that for 150 of these patients the BG or FBG levels had been above their diagnostic criteria (Section 6.1.1) at the index date. Table 6.3 traces the outcomes for these patients with undiagnosed diabetes over the next 10 years.

Table 6.3 Patients with undiagnosed type 2 DM prior to the study and their outcomes.

Patients/year	0 yr	1 yr	2 yr	3 yr	4 yr	5 yr	6 yr	7 yr	8 yr	9 yr	10 yr
N remaining in the study	150	130	116	103	89	85	80	75	70	60	59
N diagnosed	0	28	36	42	44	46	48	49	51	51	53
N remaining undiagnosed	150	102	80	61	45	39	32	26	19	9	6
% undiagnosed		78%	62%	53%	44%	44%	38%	33%	25%	13%	10%

Out of the 150 patients who already had records of BG ≥11.1 mmol/L or FBG ≥7.0 mmol/L before entering the study, only 28 would be diagnosed by the end of year 1, meaning that 78% of the undiagnosed patients would remain without a record of diagnosis with type 2 DM throughout their 1st year in the study. By the end of the 5th year, 46 out of the initial 150 patients were given a diagnosis. Since some of the patients transferred out, the % undiagnosed (44%) was calculated relative to the number of patients from the undiagnosed cohort still remaining in the study (89). At the 10th year, 10% of the cohort still remaining in the study (59) would have exited the study without ever being given a diagnosis. These estimations only accounted for patients who had been at least once referred for a BG test prior to the study start date. The true extent of cases with undiagnosed DM is unknown, since over 70% of patients in the study had no BG measurements.

Finally, the loss of follow-up in itself could be biased. This happens for instance when patients die due to complications of a condition, or they transfer out to a different location to access better treatment. Some patients may re-enter at the same or different practice, during which they might be assigned a new ID number. The fact that the impact of these inconsistencies on the quality of the data would remain unquantified, makes working with routinely collected data both more challenging and rewarding.

6.3 The models

The research task, to predict 10-year incidence of type 2 DM, was approached with survival and classification models. The *survival* models included the classical Cox PH and ML-based survival DT algorithm. The *classification* models considered were: small-data NN, LR, and NN ensembles.

In order to investigate whether or not the inclusion of BG data could improve the accuracy of the prognosis, the models were considered in two settings: with and without BG data. The predicted output was intended to be a continuous variable [0 *to* 1] that represented the probability of developing type 2 DM by the end of 10 years. The true outcome was the binary variable [0 *or* 1] corresponding to being diagnosed with type 2 DM (1) or being diagnosis-free (0) by the end of 10 years. In order to analyse the confusion matrices and corresponding sensitivity and specificity values, the continuous predicted outcome was dichotomised into a binary variable. The threshold for this conversion was set to the 75^{th} percentile, in order to account for the model sensitivity to large class imbalance in the true outcomes (Section 6.4). No dichotomising was required for the computation of concordance measures, which operated on continuous outputs.

The derivation cohort applicable for a given model differed from one model to another, due to the varying limitations and advantages of each model. In order to ensure

consistent comparisons among the NN ensemble, Cox PH, Survival DT and LR models, they were evaluated on the *same* cohort of 26653 patients, unaltered and as originally sampled.

6.3.1 Cox proportional hazards model

Designed to handle censored samples, a Cox PH model was developed with the entire derivation cohort of 53306 samples, including all of the 1585 DM, 17907 non-DM and 33814 unknown outcome records available for the model derivation. The model presented here is the prototype version of the Cox PH model developed in collaboration with the Oxford group [90] which was derived separately for men and women and included time-varying coefficients and multiple polynomial terms. Despite the difference in complexity, this prototype achieved the same (to 2 decimal places) concordance and prognostic performance as the average of the male and female benchmark Cox PH models. In accordance with the collaboration protocol [90], missing BMI values were imputed using the MICE [290] with 100 iterations and the missing BG values were imputed zero, and additional variable BG_{pres} was supplied to indicate presence or absence of BG values.

The resulting model provided the hazard of the event "diagnosis with type 2 DM" happening relative to the baseline hazard $\lambda_o(t)$, and took the following form:

$$\frac{\lambda(t)}{\lambda_o(t)} = e^{(\beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p)} \qquad eq.\ 6.1$$

where x are predictor variables and β are the estimated model parameters. The values of β and corresponding hazard ratio (HR) e^{β} for the variables x in the Cox PH models with and without the inclusion of BG are provided in Tables 6.4 and 6.5 respectively.

Table 6.4 Cox PH model without blood glucose information

Variable x	Coefficient β	Hazard ratio e^{β}	95% CI		p-value
			lower	upper	
Gender female	-0.581	0.559	0.505	0.620	<0.001
Age (years)	0.039	1.040	1.035	1.044	<0.001
Family history of DM	0.509	1.664	1.378	2.010	<0.001
CVD	0.337	1.401	1.195	1.643	<0.001
BMI (kg/m²)	0.130	1.139	1.130	1.147	<0.001
Hypertension	0.452	1.572	1.396	1.771	<0.001
Ethnicity "Asian"	0.996	2.707	1.988	3.688	<0.001
Ethnicity "Black"	0.246	1.279	0.764	2.140	0.349
Ethnicity "Mixed"	0.087	1.091	0.453	2.627	0.846
Ethnicity "Other"	1.060	2.886	1.845	4.516	<0.001
Prescribed steroids	0.279	1.321	1.061	1.646	0.013
Smoker	0.291	1.338	1.187	1.507	<0.001
Townsend score	0.064	1.066	1.049	1.083	<0.001

Table 6.5 Cox PH model with blood glucose information

Variable x	Coefficient β	Hazard ratio e^{β}	95% CI		p-value
			lower	upper	
Gender female	-0.533	0.587	0.530	0.651	<0.001
Age (years)	0.037	1.038	1.033	1.042	<0.001
Family history of DM	0.511	1.666	1.379	2.013	<0.001
CVD	0.253	1.288	1.097	1.512	0.002
BMI (kg/m²)	0.125	1.133	1.124	1.142	<0.001
Hypertension	0.393	1.481	1.311	1.674	<0.001
Ethnicity "Asian"	0.874	2.397	1.760	3.265	<0.001
Ethnicity "Black"	0.302	1.353	0.809	2.263	0.250
Ethnicity "Mixed"	0.059	1.061	0.441	2.557	0.895
Ethnicity "Other"	1.051	2.861	1.829	4.474	<0.001
Prescribed steroids	0.218	1.243	0.998	1.549	0.053
Smoker	0.318	1.375	1.220	1.549	<0.001
Townsend score	0.060	1.062	1.045	1.080	<0.001
BG recorded	-2.472	0.084	0.064	0.112	<0.001
BG level (mmol/L)	0.483	1.621	1.552	1.694	<0.001

The inclusion of BG information improved the prognostic ability of the model from C-index = 0.809 to C-index = 0.825 on the model development cohort. When validated with the independent test cohort of 26653 samples, the Cox model *without* BG performed

with C-index $= 0.817$, 73% sensitivity and 74% specificity. The Cox model *with* BG achieved C-index $= 0.832$, and was able to correctly stratify 76% of DM and 74% non-DM groups among patients with known outcomes.

According to the baseline Cox model without BG levels, the highest risk factors (HR $\geq$ 1.5) were: ethnicity "Asian" and "Other", male gender, presence of family history of DM, and history of being treated for hypertension. The control group were patients with ethnic origin "White", no family history of DM, and no history of hypertension respectively. All variables, apart from those with small representation in the cohort (ethnicity "Black" or "Mixed", use of steroids) were statistically significant.

Inclusion of BG information marginally reduced the HR of being treated for hypertension to below 1.5. The Cox model with BG confirmed the hypothesis that elevated BG levels were a high-risk factor for type 2 DM. Surprisingly, the *presence* of BG measurements was negatively associated with type 2 DM, which could be due to the information overlap of BG levels with BG presence. To examine this further, a third Cox PH model was developed to include only the binary variable for BG presence (without the corresponding BG levels). The model was largely similar to the one presented in Table 6.4, with the additional positive association (HR =1.212) of having BG recorded with the type 2 DM diagnosis (p-value <0.001), which contradicted the findings in Table 6.5, thus necessitating further investigation with models that could prove less sensitive to correlated cofactors.

Combined, these findings demonstrated that:

- Inclusion of BG measurements increased the prognostic value (C-index) of the Cox PH model from 0.809 to 0.825.

- Patients of "Asian" and "Other" ethnic origin, patients of male gender, patients with an existing family history of DM or a history of being treated for hypertension, and patients with elevated BG levels had an over 1.5 times higher risk of developing type 2 DM than the control group.
- The relative importance of the variables remained inconclusive.

6.3.2 Neural network ensemble

The NN model was stipulated by the CPRD study protocol [90] and formed a pivotal part of this work with over 1,200,000 NNs implemented and evaluated during the various exploratory and model development stages, amounting to over 138 days of simulation time alone. As a result of this extensive study, the model evolved from a multi-node NN to an ensemble of 100 two-layer NNs, where a single neuron in the hidden layer formed a 'bottle neck'.

The model was trained with 1585 DM and 17907 non-DM examples, and was validated on an independent cohort of 26653 patients. With the ratio of minority (DM outcome) to majority (non-DM outcome) examples being approximately 1:11, the model suffered from vast class imbalance and initially failed to learn minority class associations. The class imbalance problem was later addressed by combining ensemble learning with majority undersampling as follows:

a) Divide the majority class into 10 non-overlapping subsets
b) Use one of the majority class subsets, plus all minority class samples to train 10 individual NNs
c) Repeat (b) for the remaining 9 majority class subsets to produce 100 NNs
d) Combine the NNs into an ensemble by averaging

It is important to emphasise that the above balancing strategy was only applied to the model derivation cohort, whilst the 26653 validation cohort samples, with which NN ensemble was subsequently tested, retained its original class imbalance.

Also considered was minority oversampling with ADASYN, where over 17000 synthetic DM samples were created to match the number of non-DM samples; however, no increase in performance over the proposed strategy with the majority undersampling was observed.

The individual NNs in the ensemble were trained with a scaled conjugate gradient (SCG) backpropagation algorithm (Appendix A.2), which offered higher robustness, albeit at slower computation speed than the Levenberg-Marquardt backpropagation implemented with multiple runs strategy in Chapters 3 and 4. The two-layer architecture using a *tan-sigmoid* function in the hidden layer and *log-sigmoid* function in the output layer was developed to provide as much separation between low-DM-risk and high-DM-risk patients in the interval between 0 and 1. The cost function used was *cross entropy* between predicted risk and known binary outcome. For the purpose of computing confusion matrices, and the corresponding sensitivity and specificity values, a binary version of the predicted outcome was generated from the continuous risk value using a threshold set at 0.5. This was different from the 75th percentile threshold used with Cox PH, since each NN in the ensemble was trained with balanced data.

Four core *data models* representing various degrees of BG completeness were investigated:

1) without any BG levels ("no BG")
2) with BG presence and aggregate BG levels imputed zero if missing ("BG new")
3) with aggregate BG levels imputed with MICE if missing ("BG imputed")

4) with random BG and FBG imputed with MICE if missing ("BG and FBG imputed")

Their corresponding performance on the 26653 independent test samples is provided in Table 6.6. The classification measures, i.e. sensitivity Sn, specificity Sp, and balanced accuracy $C_{balanced}$, were evaluated on 9810 samples with known outcomes.

Table 6.6 NN ensemble performance: concordance and classification measures

NN ensemble models	C-index	95% CI		Sp	Sp	C_{balanced}
		lower	upper			
no BG	0.829	0.816	0.842	0.629	0.835	0.732
BG new	0.847	0.834	0.860	0.722	0.807	0.765
BG imputed	0.901	0.891	0.911	0.798	0.866	0.832
BG and FBG imputed	0.929	0.920	0.938	0.855	0.887	0.871

The first two models ("no BG" and "BG new") correspond to the two Cox PH models described in Section 6.3.1. As with Cox PH, the inclusion of BG values improved the NN ensemble performance on the validation cohort from C-index = 0.829 to 0.847. Unlike Cox PH, ensemble learning makes NNs largely a "black-box" system with no direct interpretation of variable importance [21,291].

The two latter models revealed a substantial increase in C-index performance: 0.901 for "BG imputed" and 0.929 for "BG and FBG imputed". In other words, the NN ensemble, when using information on both BG and FBG values, was able to predict the 10-year type 2 DM outcome with nearly 93% accuracy and 12% more reliably than the model without BG. This finding, initially thrilling, was dismissed following further consideration. Firstly, the imputed models relied on the vast proportion of the BG (nearly 80% missing) and FBG (over 93% missing) values being synthetically generated by MICE. Secondly, since the MICE imputation model was developed separately for the DM and non-DM groups, it was possible that the NN was able to decode the DM and non-DM groups from the imputed BG and FBG variables. Finally, when re-evaluated on only non-imputed test

samples ($n = 625$), the performance of the model, at C-index = 0.796, proved inferior than that of "no BG" model. This decrease in performance was due to false positive samples: i.e. the sensitivity remained high (83%), but specificity fell to 56%, resulting in 70% balanced accuracy. Combined, these outcomes indicated the deceptive effectiveness of imputed data, and precluded the use of MICE with NNs on data where the proportion of missing values is as high as 80-90%.

To summarise, it has been demonstrated that:

- The proposed approach of combining majority undersampling with ensemble learning offered a simple and effective solution, of which the performance exceeded that of state-of-the-art synthetic minority oversampling techniques.
- The prognostic value of the NN ensemble was improved with the inclusion of BG measurements, from C-index of 0.829 to 0.847.
- The dramatic 12% improvement to C-index = 0.929 achieved by the NN ensemble with synthetically imputed BG and FBG was deemed unreasonable - a finding that revealed the pitfalls of applying multiple imputation to data with over 80% missing values.

6.3.3 Small-data neural network

One frequently overlooked aspect of modelling with sparse samples is that of a complete-case scenario. The research question investigated in this section was whether or not a well-generalising NN could be developed with a small, but high-quality data subset, and if so, what would be its performance on a validation cohort, given the complete-case scenario.

Out of the 79959-patient dataset, there were only 1415 patients (918 in model derivation and 497 in validation cohorts), for whom BG, FBG and BMI measurements were recorded at least once during 5 years prior to the patient entering the study. In the model derivation cohort, 521 patients were lost to follow-up, effectively reducing the dataset available for supervised learning from over 19000 to 397 samples (318 non-DM and 79 DM examples). Further class balancing would yield a training dataset of merely 158 samples. With less than 7 event observations per predictor variable, the task of predicting the long-term incidence of type 2 DM becomes a *small-data* problem.

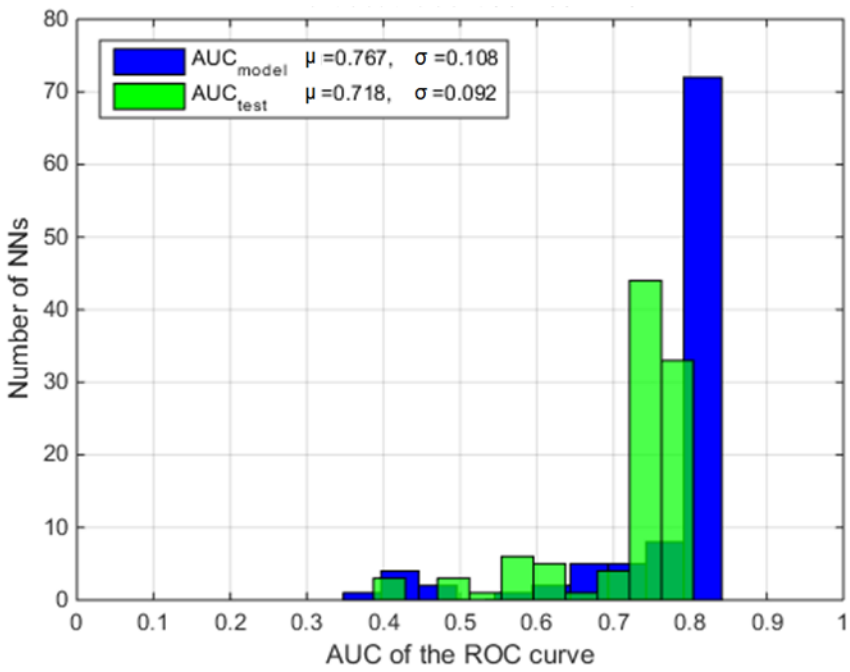


Figure 6.5 Small-data performance over a run of 100 NNs on test and model samples (η=1, ω=18)

The development of a NN model for type 2 DM under the small-data conditions followed the methodological framework developed in Chapter 3. The base NN configuration was similar to that in the NN ensemble discussed in Section 6.3.2. A two-layer backpropagation NN with 12 inputs and 1 output was trained with a SCG algorithm in order to optimise the cross entropy between predicted and output values. Early stopping

was applied to reduce overtraining. The classification performance of the NNs in each run was assessed by the median AUC (defined in Appendix B) on the model derivation cohort and, separately, on independent tests. The *method of multiple runs* made it possible to evaluate and optimise the NN design parameters despite the output volatility due to small data (Figure 6.5).

The hidden layer size η and the training duration (as controlled by the early stopping criterion ω) were optimised in an iterative simulation involving 280 runs of 100 NNs. The effect of ω varying from 5 to 20 was highest at $\omega = 18$, but did not prove statistically significant (pairwise p>0.05) in comparison to other values of ω.

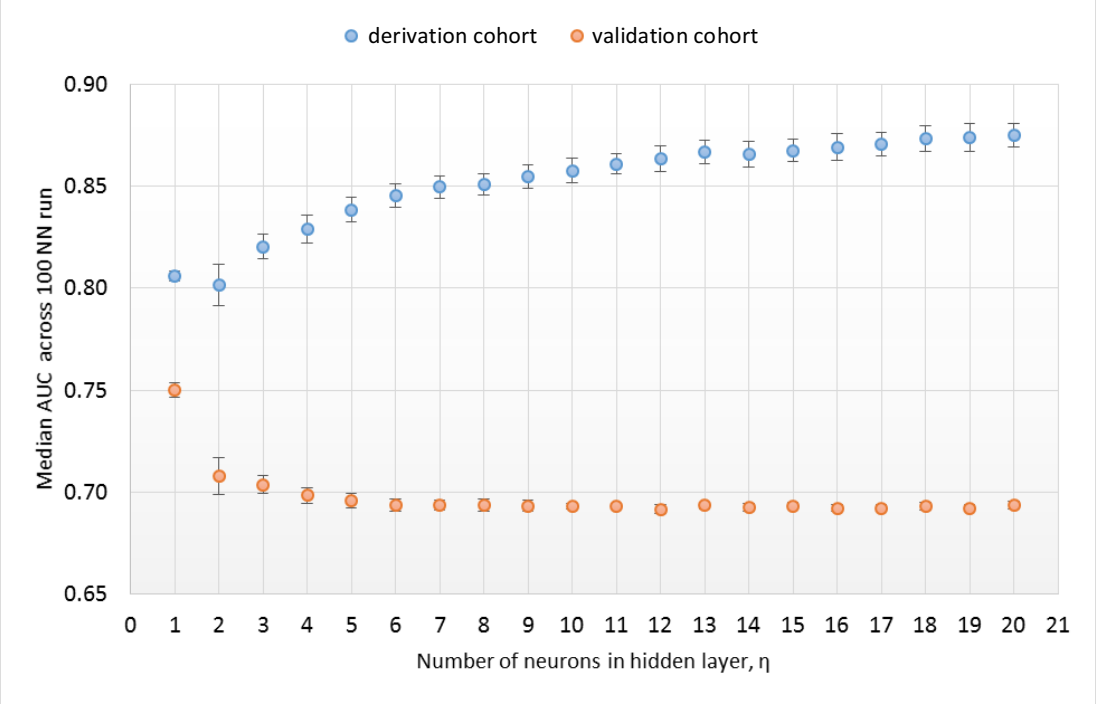


Figure 6.6 The small-data NN design optimisation: effect of the hidden layer size on the NN performance

The hidden layer size varying from $\eta = 1$ to $\eta = 20$ had a more pronounced effect on the NN performance (Figure 6.6). For the training cohort, the median AUC across the run increased monotonously with the increasing η, since larger networks were able to learn the patterns with greater ease in the model cohort. Contrary, the AUC for the validation

cohort was highest at $\eta = 1$ and decreased with increasing η, indicating that the improvement observed in the model performance was due to overtraining. The single hidden layer forming a "bottleneck" was significantly more effective than any other hidden layer size considered (pairwise $p < 0.01$, Wilcoxon rank sum test).

Out of the 100 small-data NNs in the optimal $\eta = 1$, $\omega = 18$ run (Figure 6.5), the best performing model achieved AUC = 0.834 on derivation cohort and AUC=0.804 on the 207 test samples with known outcome (Figure 6.7).

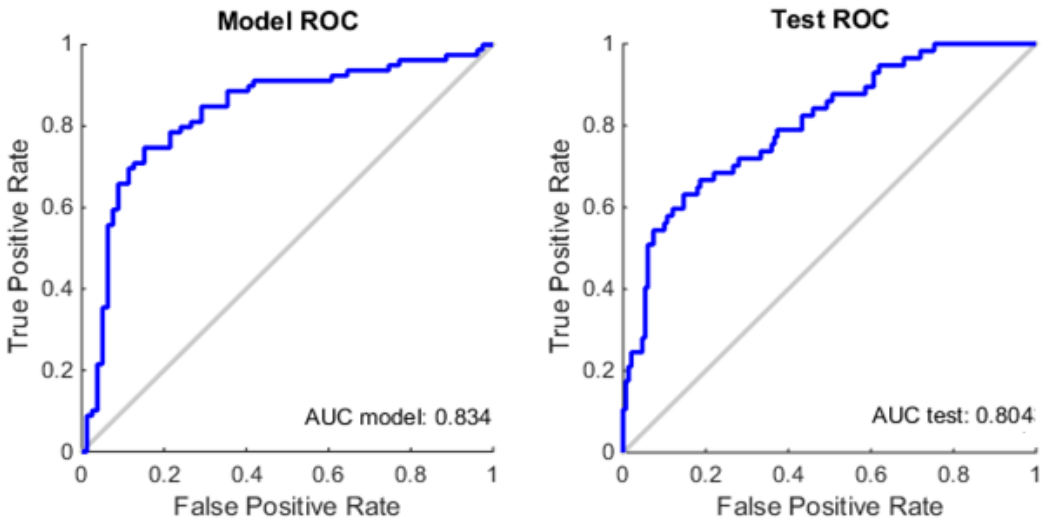


Figure 6.7 The small-data NN: Receiver operating characteristic (ROC) curves for model derivation and validation (test) cohorts

With $\eta = 1$, the input weights w_I became a 12 x 1 vector, and the hidden layer bias $b^{(1)}$ became a scalar. The resulting NN classifier evaluated score y of whether or not a patient with baseline indicators x would develop type 2 DM in 10 years. Its output equation could be written as:

$$y = logsig(tansig(xw_I + b^{(1)})w_L + b^{(2)}) \qquad eq.\ 6.2$$

Substituting the *logsig* and *tansig* functions (Appendix A.1) in *eq*. 6.2, the output takes the form of:

$$y = \frac{1}{1 + e^{-b^{(2)} - w_L \frac{e^{xw_I + b^{(1)}} - e^{-\left(xw_I + b^{(1)}\right)}}{e^{xw_I + b^{(1)}} + e^{-\left(xw_I + b^{(1)}\right)}}}} \qquad eq.\ 6.3$$

where parameters w_I, w_L, $b^{(1)}$, and $b^{(2)}$ are determined during the NN training. Unlike in the NN ensemble model, in this stand-alone NN the weights could be traced from each input variable to the bottle-neck hidden layer neuron, thus helping to reveal a partial indication of how a given variable affected the predicted output. The input weights w_I were used as a measure of *relative* variable importance in the NN predictions (Figure 6.8). Unsurprisingly, the fasting BG levels had by far the strongest prognostic value in the NN model, adding incentive to the inclusion of BG information for any future type 2 DM prediction model.

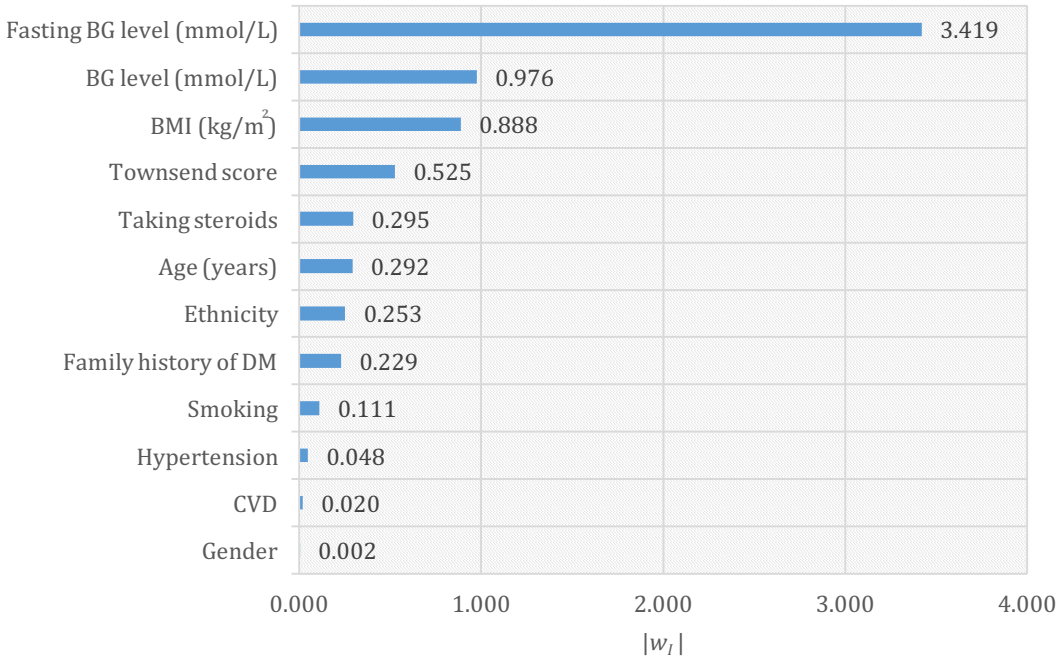


Figure 6.8 The not-so-black-box NN: relative variable importance by the absolute values of input weights

When evaluated on all 497 test samples available for complete-case scenario (both for missing and known outcome), the small-data NN, trained with only 158 samples, was able to achieve concordance of C-index $= 0.783$, which was on par with the performance

of the far more advanced NN ensemble developed with over 19,000 samples and the Cox model built on over 53,000 samples.

To summarise this complete-case investigation:

- The methodological framework developed in Chapter 3 was successfully applied to design a 78% accurate NN classifier with less than 7 event observations per predictor variable.
- Despite the selection bias associated with the complete-case scenario, the model was able to generalise on 497 independent test samples, including patients for whom the 10-year outcome was unknown.
- In an extensive simulation involving 28000 NNs, a 1-neuron hidden layer "bottleneck" design proved optimal for the given task.
- The fasting BG level, followed by random BG level, BMI and Townsend deprivation score had the strongest predictive value in the 1415 patients included in the complete-case scenario.

6.3.4 Logistic regression

The development of the LR model for type 2 DM prediction was motivated by Section 6.3.3, where it was discovered that a single *tan-sigmoid* neuron in the hidden layer yielded the best NN model fit. If we view the function implemented by the hidden neuron $f_h(x) = \frac{e^{xw_I + b^{(1)}} - e^{-\left(xw_I + b^{(1)}\right)}}{e^{xw_I + b^{(1)}} + e^{-\left(xw_I + b^{(1)}\right)}}$ as merely *an input transformation*, then $eq.$ 6.4 for the NN output becomes:

$$y = \frac{1}{1 + e^{-b^{(2)} - w_L f_h(x)}} \qquad eq.\ 6.4$$

where $b^{(2)}$ and w_L are scalars. It could be noted that *eq.* 6.4 is similar in form to the output of a standard LR model:

$$y = \frac{1}{1 + e^{-logit(x)}} \qquad eq.\ 6.5$$

$$logit(x) = \beta_o + \beta_1 x_1 + \beta_1 x_1 + \cdots + \beta_p x_p \qquad eq.\ 6.6$$

Noting this resemblance, it was decided to explore whether the 10-year incidence of type 2 DM could be successfully modelled with a conventional LR. The LR model was developed with 1585 DM and 17907 non-DM examples, and validated on an independent cohort of 26653 patients. The DM and non-DM classes were balanced by SMOTE. As with Cox PH and NN ensemble models, two scenarios were considered: one without the inclusion of BG information and one with the inclusion of any available BG level and the corresponding presence flag. The set of β parameters of the resulting two LR models are provided in Tables 6.7 and 6.8 respectively.

Table 6.7 LR model without blood glucose information

Variable x	Coefficient β	Odds ratio e^{β}	95% CI		p-value
			lower	upper	
Intercept	-6.706	0.001	0.001	0.002	<0.001
Gender female	-0.722	0.486	0.410	0.575	<0.001
Age (years)	0.044	1.045	1.038	1.052	<0.001
Family history of DM	1.039	2.827	1.911	4.236	<0.001
CVD	0.454	1.574	1.138	2.199	0.007
BMI (kg/m^2)	0.154	1.166	1.147	1.186	<0.001
Hypertension	0.458	1.581	1.272	1.969	<0.001
Ethnicity “Asian”	3.043	20.971	7.137	89.969	<0.001
Ethnicity “Black”	1.269	3.557	1.164	13.375	0.037
Ethnicity “Mixed”	0.048	1.049	0.226	4.750	0.950
Ethnicity “Other”	2.665	14.361	3.792	94.494	<0.001
Prescribed steroids	0.424	1.528	1.010	2.335	0.047
Smoker	0.577	1.780	1.452	2.185	<0.001
Townsend score	0.082	1.086	1.054	1.119	<0.001

Table 6.8 LR model with blood glucose information

Variable x	Coefficient β	Odds ratio e^{β}	95% CI		p-value
			lower	upper	
Intercept	-6.428	0.002	0.001	0.003	<0.001
Gender female	-0.693	0.500	0.421	0.594	<0.001
Age (years)	0.041	1.042	1.035	1.050	<0.001
Family history of DM	1.076	2.933	1.965	4.433	<0.001
CVD	0.352	1.422	1.014	2.010	0.044
BMI (kg/m^2)	0.147	1.158	1.139	1.178	<0.001
Hypertension	0.401	1.494	1.188	1.880	<0.001
Ethnicity "Asian"	3.030	20.702	7.004	89.038	<0.001
Ethnicity "Black"	1.310	3.704	1.209	13.948	0.032
Ethnicity "Mixed"	0.077	1.080	0.241	4.743	0.918
Ethnicity "Other"	2.633	13.919	3.683	91.528	<0.001
Prescribed steroids	0.440	1.552	1.020	2.384	0.042
Smoker	0.563	1.757	1.428	2.164	<0.001
Townsend score	0.084	1.088	1.055	1.122	<0.001
BG recorded	-3.864	0.021	0.008	0.053	<0.001
BG level (mmol/L)	0.742	2.101	1.777	2.506	<0.001

The inclusion of BG information improved the prognostic ability of the LR model by approximately 2%, from C-index = 0.810 to C-index = 0.827, on the validation cohort. The LR model *without* BG was able to correctly stratify 71% of DM and 76% non-DM groups. The LR model *with* BG achieved 74% sensitivity and 77% specificity among patients with known outcome.

The BG model parameters with the highest odds of type 2 DM at 10 years were largely similar to those established by Cox PH in Section 6.3.1. Patients of "Asian" and "Other" ethnic origin, patients of male gender, patients with an existing family history of DM, and patients with elevated BG levels remained a high-risk group with over 2 times odds of developing type 2 DM at 10 years. Similarly to the Cox PH model, the *presence* of BG measurements in LR model was negatively associated with the outcome.

To summarise, these findings demonstrated that a standard LR classifier was marginally inferior in performance to the more complex NN ensemble, which benefited from tan-sigmoid transformation in the hidden layer. The LR model confirmed the associations previously established by the Cox PH model on the derivation cohort.

6.3.5 Survival decision tree

A survival DT model offered a mechanism for dealing with censored outcomes and missing covariates, whilst also producing a concise graphical representation of high-risk groups. The DT was developed with all of the 1585 DM, 17907 non-DM and 33814 unknown outcome records available for the model derivation. No imputation of BMI or BG was required. Two survival DT models were considered: one without the inclusion of BG information (Figure 6.9) and one with the BG values and BG presence indicator (Figure 6.10).

The survival DTs implemented in this work were based on the local full likelihood tree model of LeBlanc & Crowley [292]. The DT partitioned the covariate space into subsets of patients based on the log rank criterion, so that every new partition increased the homogeneity of the observations within each patient group. The proportion of patients at every node who were diagnosed with type 2 DM was then evaluated and compared with that at the root node. This *relative event rate* r indicated the hazard of developing type 2 DM; the relative event rate at the root was r = 1, which is equivalent to the baseline hazard in Cox PH. The model output y was expressed as r mapped between 0 and 1, i.e. $r = \frac{r - \min(r)}{\max(r) - \min(r)}$. In order to assess the classification measurements, the output of the DT was dichotomised as follows:

$$y_{binary} = \begin{cases} 1, if\ r > 1 \\ 0, if\ r \leq 1 \end{cases} \qquad eq.\ 6.7$$

In order to prevent overfitting, the minimum parent size was set to $minsplit = 100$. A pruning complexity parameter of $cp = 0.003$ was specified to control DT growth: any split that did not improve the fit by a factor of 0.003 was not attempted.

Surrogate splits were constructed for each node, which allowed the handling of variables with missing values. If an observation missed the primary and all possible surrogate splits, then the DT sent the observation in the majority direction. As a result of surrogate splits, a variable could appear in the DT multiple times both as primary and surrogate. The relative *variable importance scores* were computed from the combined node purity for every split (surrogate or otherwise), in which the variable in question had featured.

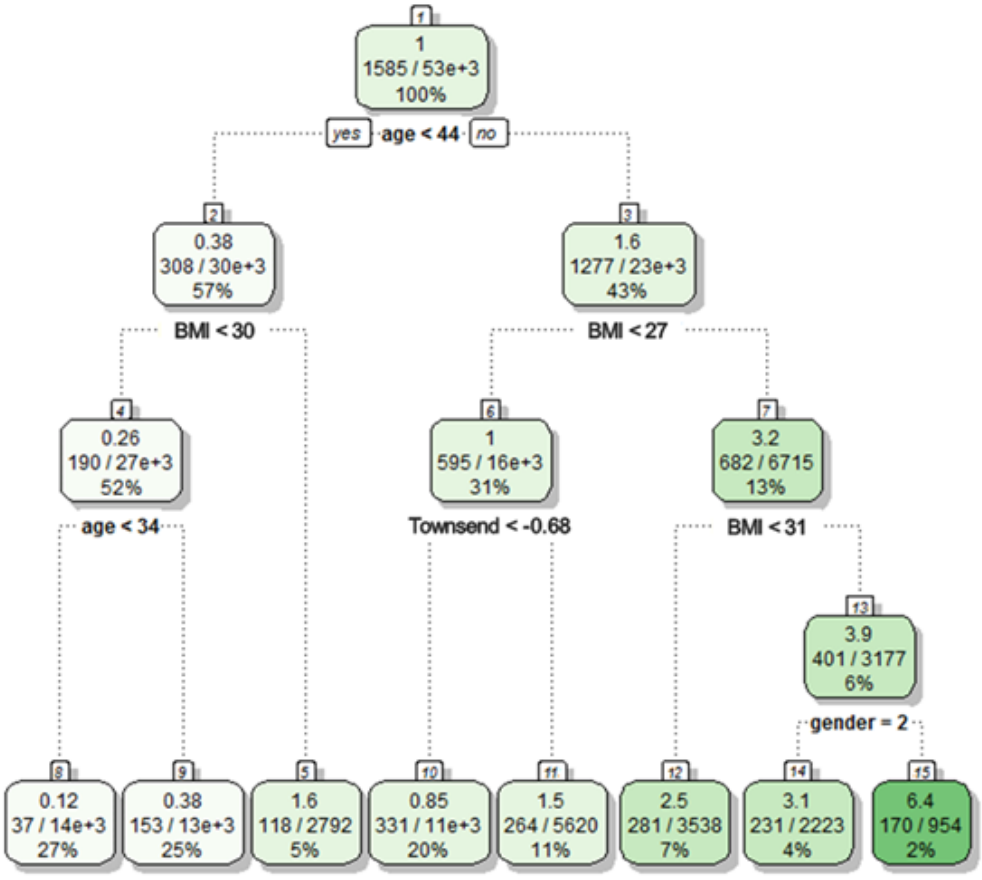


Figure 6.9 Survival DT modelled without the inclusion of BG (missing values of BG and BMI left unaltered).

The DT in Figure 6.9 was modelled on the derivation cohort without BG values. It consisted of 7 branch nodes and 8 terminal nodes, numbered 1 to 15. The box under each node provides the value of r, the number of events/total number of samples passing through the node, and the % of the derivation cohort size. The colour intensity is associated with the higher relative likelihood of being diagnosed with type 2 DM.

Without the BG information, the survival DT achieved a C-index of 0.786 on the validation cohort. The DT model was able to correctly stratify 64% of DM and 70% of non-DM patients. The following variables (in decreasing order of importance) were used by the DT in primary and surrogate splits: age, BMI, hypertension, gender, CVD, Townsend score and ethnicity.

Of a particular interest was the DT modelled *with* BG shown in Figure 6.10, where all primary splits were based on continuous variables (age, BMI and BG). In the DT's surrogate splits, the following additional variables were used (in decreasing order of importance): hypertension, presence of BG, CVD, steroid use, Townsend score and ethnicity. This DT stratified patients in the derivation cohort into 9 unequally-sized risk groups, which are represented by the terminal nodes.

The 10-year risk of type 2 DM was highest in patients at nodes 15, 17, and 11, comprising:

- Individuals 44 years or older with BMI in the overweight range (≥ 26 kg/m^2) (r = 8.3, node 15), particularly if their BG level ≥ 6.4 mmol/L (r = 12, node 17);
- Individuals younger than 44 years, but who were both obese (BMI ≥ 30 kg/m^2) and had elevated BG levels ≥ 6.4 mmol/L (r = 7, node 11).

Patients younger than 44 years of age, with BMI < 30 kg/m^2, were the least likely to develop type 2 DM (r = 0.26, node 4). This group constituted 52% of the patients in the derivation cohort. The second largest stratum (r = 1, node 12), comprising 35% of the

cohort, was less conclusive: its estimated event rate was identical to that of the whole cohort. The individuals that fell into this stratum were 44 years or older, with normal BG (<5.8 mmol/L) and BMI below obese range (<30 mmol/L). With 644 events assigned for the baseline hazard, this stratum was the highest producer of *false negatives*, signifying that patients in this cohort should be considered with additional care. As a result of the false negatives, the classification accuracy of the DT model *with* BG suffered from the imbalance between the sensitivity of 45% and the specificity of 88%, indicating the need for more granularity in the terminal node, in particular at node 12.

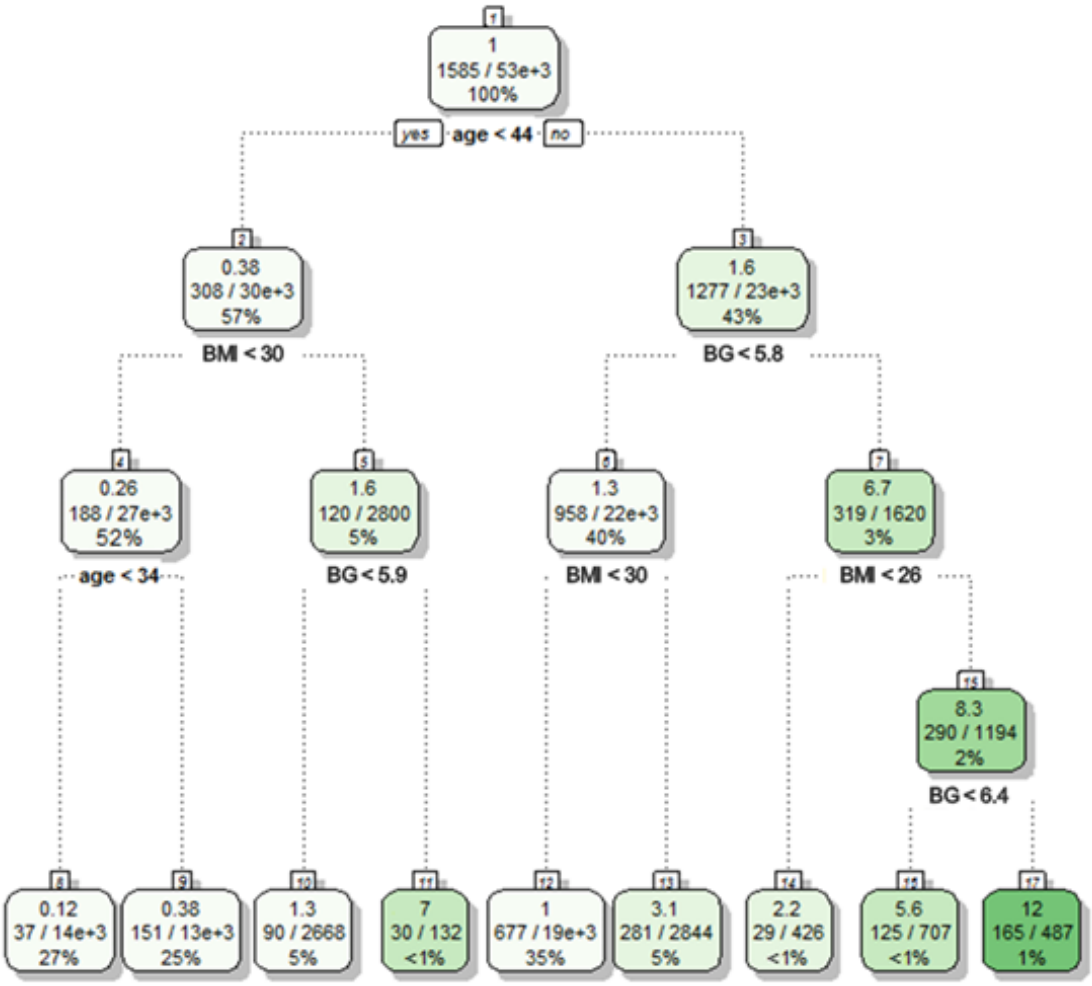


Figure 6.10 Survival DT modelled with the inclusion of BG (missing values of BG and BMI left unaltered).

Kaplan-Meyer survival curves in Figure 6.11 visualise the difference in prognosis in the patient groups represented by each terminal node for the DT, with and without BG. Inclusion of BG information improved the prognostic value of the DT by nearly 5% to a C-index of 0.824 on the validation cohort.

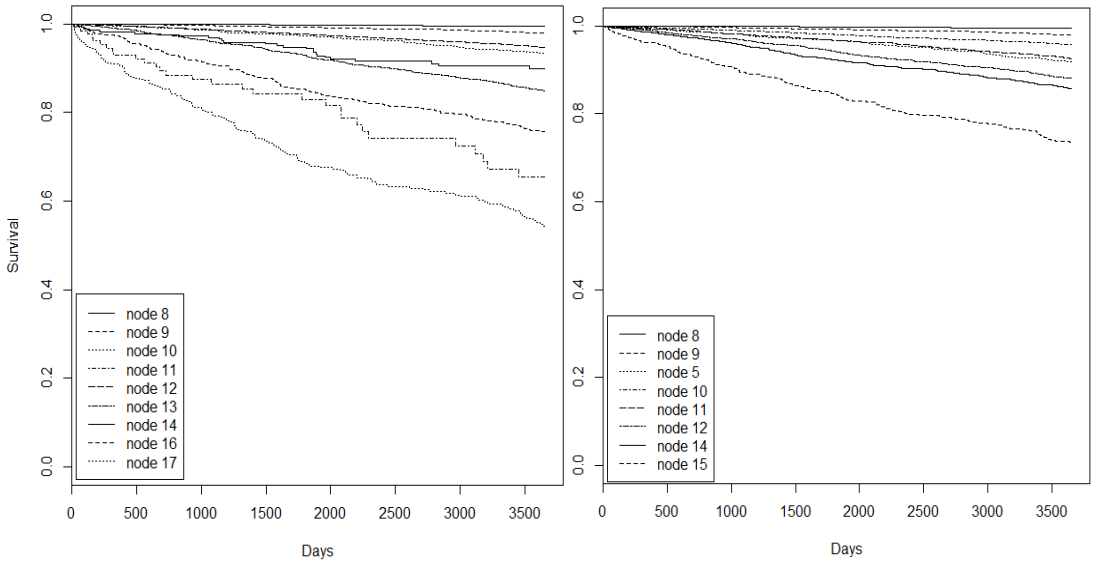


Figure 6.11 Kaplan-Meyer curves for the terminal nodes of the DT with BG (left) and without BG (right)

The variable importance scores for the two DT models in Figure 6.12 demonstrate that *age* is the most important parameter for both models, followed by BG level, BMI, treatment for hypertension, presence of BG measurements, and CVD. For the DT model without BG, gender and Townsend scores were also important factors. Family history of DM was not used in either survival DT models (neither as a primary nor secondary split). Ethnicity played a smaller role than in the Cox PH model since in the survival DTs, the small proportion of minority ethnics meant that their relative contribution against the baseline split demonstrated only marginal improvements.

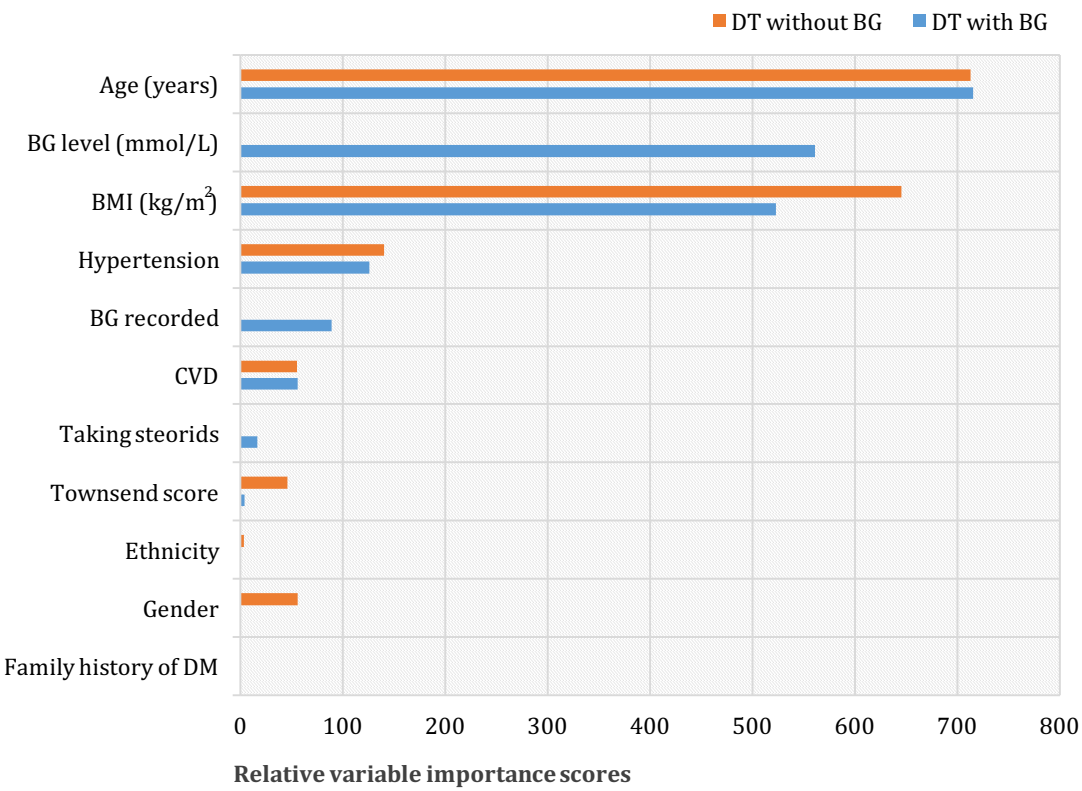


Figure 6.12 Variable importance scores for DT with BG (blue) and DT without BG (orange)

To summarise:

- The DT models were able to successfully stratify the 10-year risk of type 2 DM without relying on missing data imputation.
- The inclusion of BG information led to a significant improvement in the prognostic value, which increased by nearly 5%, from a C-index of 0.786 to a C-index of 0.824, on the validation cohort.
- Despite the large proportion of missing values in BMI and BG levels, these two variables, together with the patient's age, exhibited the highest relative importance in the DT model, emphasising their prognostic value in the prediction of type 2 DM.

6.4 Model performance and limitations

Using Cox PH as the benchmark model, the prognostic performances of the NN ensemble, small-data NN, LR and survival DT models were evaluated using Harrell's C-index and Royston's and Sauerbrei's D and R_2^D scores (Table 6.9). For validation cohort samples where the outcome was known, standard classification measures, such as specificity Sp, sensitivity Sn, and balanced accuracy $C_{balanced}$ were also assessed (Table 6.10).

Table 6.9 Comparison of model performance: Harrell's C and Royston's D measures of discrimination

	Harrell's concordance index				**Royston and Sauerbrei's D factor**				
	C-index	95% CI		n pairs	D	95% CI		R_D^2	n
		lower	upper			lower	upper		
1. Models without BG									
Cox PH	0.817	0.803	0.831	28004520	7.08	6.32	7.92	0.628	26653
NN ensemble	0.829	0.816	0.842	28004244	7.28	6.52	8.13	0.635	26653
Small-data NN	0.625	0.550	0.699	33718	2.39	1.54	3.71	0.363	497
Logistic regression	0.810	0.796	0.825	28004520	6.84	6.09	7.67	0.620	26653
Survival DT	0.786	0.770	0.803	24478846	4.86	4.32	5.48	0.537	26653
2. Models with BG									
Cox PH	0.832	0.819	0.846	28004522	8.44	7.53	9.45	0.668	26653
NN ensemble	0.847	0.834	0.860	28004338	9.14	8.17	10.23	0.686	26653
Small-data NN	0.783	0.724	0.842	33720	6.29	4.01	9.87	0.600	497
Logistic regression	0.827	0.813	0.841	28004522	7.94	7.08	8.89	0.655	26653
Survival DT	0.824	0.807	0.842	22185650	5.92	5.26	6.66	0.586	26653

Harrell's C-index measured concordance, which is defined in Appendix B as the proportion of all comparable pairs of patients (n pairs) where patients with longer survival time are assigned a lower risk. It is also interpreted as AUC for right-censored data [293]. The most concordant models were: NN ensemble with BG (0.847), Cox PH with BG (0.832) and NN ensemble without BG (0.829). The inclusion of patient's BG information improved all six models, but this effect was most pronounced in the models that did not leverage imputation of missing variables, i.e. small-data NN (25% increase

in C-index) and survival DT (5% increase in C-index). The improvement in C-index for Cox PH, NN ensemble and LR was only 2%.

As discussed in Appendix B, Royston's and Sauerbrei's D divides the distribution of the patient prognostic indices into two equally-sized risk groups at the median value and compares their relative hazard. This property allows the D score to be interpreted as the model's overall log hazard ratio [293]. Another way to reason about the D score is through its R_D^2 transformation, which provides a measure of prognostic separation in the interval bound between 0 and 1. The highest D was achieved by models with BG: NN ensemble (9.14), Cox PH (8.44) and LR (7.94). Notably, the largest disagreement between C-index and D score was in the survival DT model. This stems from the differences in the output distributions produced by each prognostic model (Figures 6.13 and 6.14). Since type 2 DM was a rare event, it meant that the use of medians in Royston's and Sauerbrei's D score was less appropriate for some models than others.

The outputs y produced by each model should represent *prognostic indices* related to the probability of being diagnosed with type 2 DM, and hence were designed to lie in the common interval between 0 and 1. To achieve this, the raw outputs y_0 of the Cox PH and the survival DT models (which have no upper bound) were scaled as follows:

$$y = \frac{y_o - \min(y_o)}{\max(y_0) - \min(y_0)} \qquad eq.\ 6.8$$

Such transformation preserves the prognostic separation and does not disturb the model performance measures. The outputs of the NN and LR models were designed from the onset to be between 0 and 1.

Four histograms in Figure 6.13 correspond to the prognostic indices y predicted by the Cox PH, ensemble NN[4], LR, and survival DT models *with* BG, where the sum of frequencies across the 20 bins corresponds to 26653 validation cohort samples. To compensate for loss of detail due to the overlaying histograms, the distributions of y are also visualised as smooth density curves, approximated by a kernel density function [294] and scaled to the interval between 0 and 1 for consistent representation (Figure 6.14). Unlike the continuous output of the Cox PH, ensemble NN and LR models, the survival DT output comprised a finite number of discrete values, corresponding to the number of terminal nodes in the tree (8 for the model *without* BG and 9 for the one *with* BG).

For the classification measures in Table 6.10, instead of using medians for dividing the prognostic indices into DM and non-DM groups, the threshold was evaluated for each model separately by accounting how well it handled class imbalance. For instance, in small-data NNs, the class imbalance was less pronounced and binary classification was achieved by using a rigid threshold of 0.5 for assigning the predictions to DM and non-DM classes. For Cox PH, NN ensemble and LR models, the 75th percentile value was a more appropriate threshold for dichotomising the prognostic indices into non-equally sized DM and non-DM groups. For surrogate DT, the threshold of 1 was applied prior to the min-max mapping to preserve interpretability of the relative event rate ρ.

The threshold for dichotomising the prognostic index could be tuned to allow for desired Sn, but at the expense of lowering Sp, and vice versa. In practice, the threshold value of a prognostic system would depend on what false positive or false negative rates an individual health provider is able to accept, and requires a careful consideration of the morbidity of the disease and the costs of its diagnosis.

[4] The small-data NN was excluded from this analysis to avoid misrepresentation of the 26653-sample validation cohort with its subset of 497 complete records.

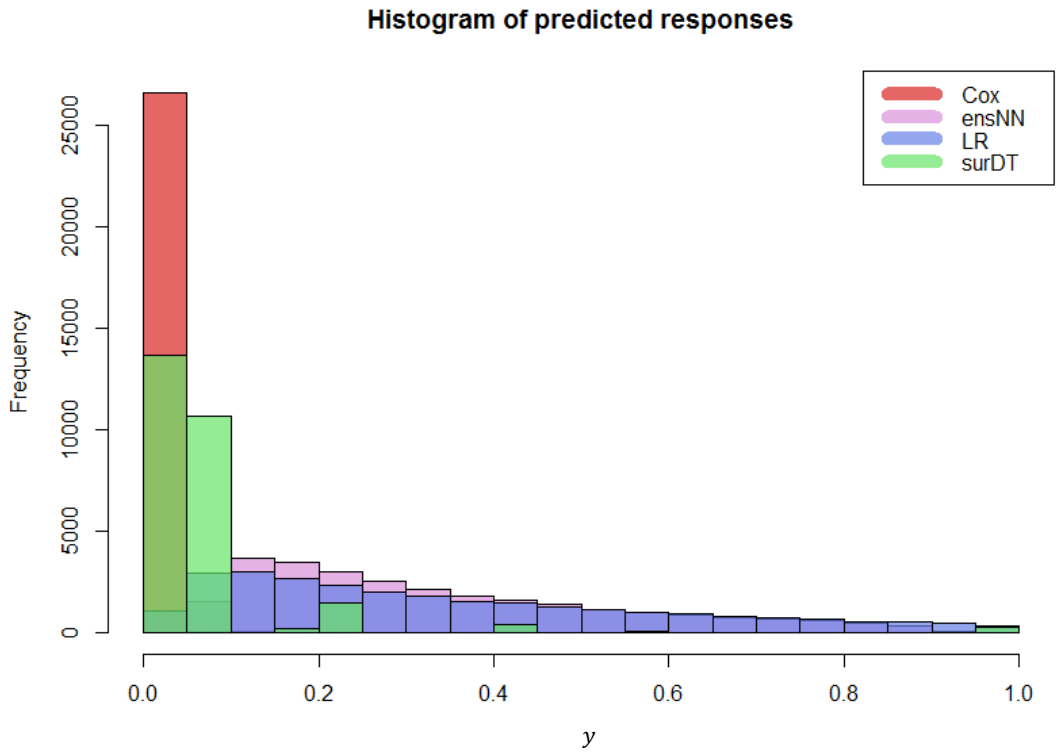


Figure 6.13 Distribution of responses predicted by Cox PH, NN ensemble, LR and survival DT models with BG.

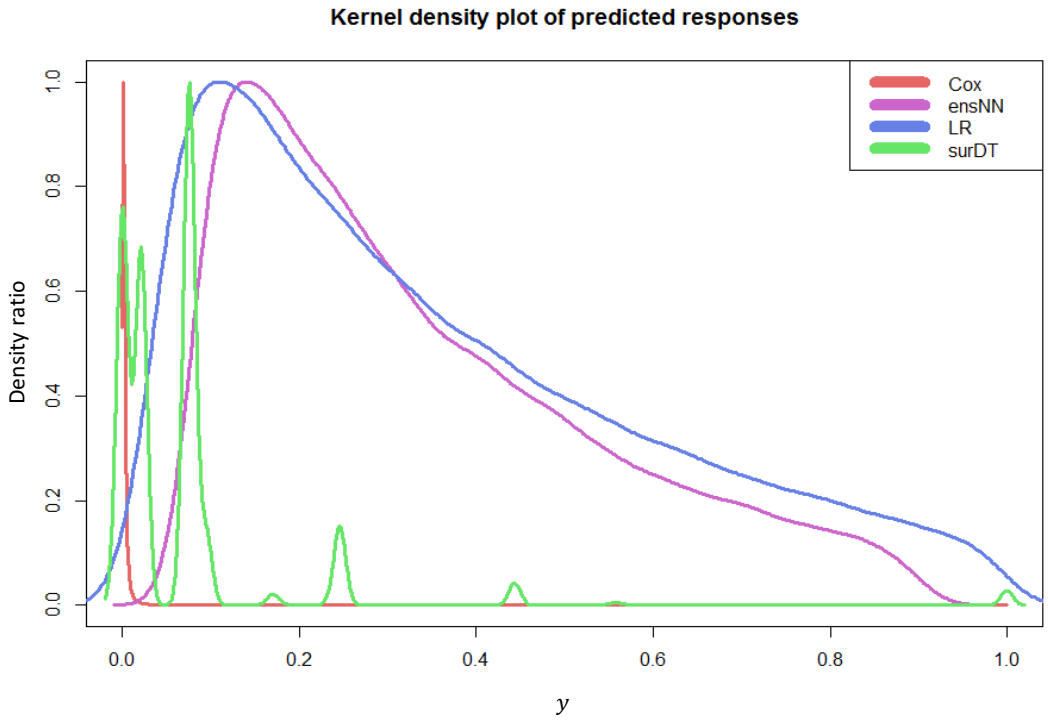


Figure 6.14 Kernel density curve of the responses predicted by Cox PH, NN ensemble, LR and survival DT models with BG.

Balanced accuracy $C_{balanced}$ (Table 6.10) assessed the number of correctly classified DM and non-DM patients from known 10-year outcomes, and represented a more conservative measure of prognostic discrimination than Harrell's C-index. When no benefit of the doubt was given to the patients who transferred out of the study before the 10-year period, all models performed equally with or without BG. The exception was small-data NN without BG information, which performed marginally better than a random coin toss. The remarkable agreement in performance among the Cox PH, LR, NN ensemble and small-data NN models with BG indicates that $C_{balanced}$ is a reflection of the quality of the *data*, rather than a property of the model.

Table 6.10 Comparison of model classification performance on samples with observed 10-year outcome

	Classification measures					
	Sn	Sp	PPV	NPV	$C_{balanced}$	n (known outcome)
1. Models without BG						
Cox PH	73%	74%	0.205	0.968	74%	9810
NN ensemble	74%	76%	0.219	0.970	75%	9810
Small-data NN	35%	79%	0.392	0.763	57%	207
Logistic regression	71%	76%	0.217	0.966	74%	9810
Survival DT	64%	70%	0.164	0.955	67%	9810
2. Models with BG						
Cox PH	76%	74%	0.214	0.971	75%	9810
NN ensemble	78%	77%	0.235	0.974	77%	9810
Small-data NN	67%	81%	0.576	0.865	74%	207
Logistic regression	74%	77%	0.228	0.970	76%	9810
Survival DT	45%	88%	0.256	0.946	67%	9810

The power of small, but high-quality data sample is further exemplified by the small-data NN. Developed with only 158 complete-case samples and suffering from a considerable exclusion bias, this NN model performed with $C_{balanced}$ equivalent to that of the NN ensemble developed with over 19,000 samples and the Cox model built on over 53,000 samples.

The NN ensemble model demonstrated its competitive performance on censored observations despite not being specifically trained to handle them. The NN ensemble model was designed with 1585 DM and 17907 non-DM examples; 33814 records with unknown outcomes were not suitable for supervised learning. The drastic reduction in the number of samples useable for model derivation – coupled with the potential selection bias in excluding those samples – was expected to negatively affect the NN performance in comparison to the Cox PH model derived with a dataset 1.7 times larger. Yet, when validated on the same independent cohort of 26653 patients, the NN ensemble model marginally outperformed Cox PH.

Notably, the survival DT exhibited poor sensitivity Sn, which *decreased* with the inclusion of BG information. As explained in Section 6.3.5, this artefact was due to an inability to differentiate, based on the available data, between DM and non-DM outcomes in one particular group of patients (node 12 in Figure 6.10): individuals 44 years or older, with normal BG (<5.8 mmol/L) and BMI below obese range (<30 mmol/L). Patients in this group accounted for 345 out of the 454 false negatives produced by the DT on the validation cohort. Should additional baseline indicators be available to explain the variance in outcome in this specific group, the sensitivity of the survival DT model could reach the theoretical maximum of $Sn = 86.3\%$[5] with the existing structure. The ability of the survival DT to pin-point not only the groups of patients with the highest hazard, but also the patients whose data require further collection is a valuable asset, not found in any other model.

[5] The theoretical maximum for the survival DT in Figure 6.10 was calculated by correcting for the node 12 artefact, while retaining the existing hierarchy of all the remaining nodes. If all 345 of the false negatives (FN) produced in node 12 were eliminated with hypothetical new baseline indicators, the overall number of FN predictions made by the DT would decrease to 454-345=109. This represents 13.2% of the 828 total type 2 DM patients in the validation cohort and corresponds to a model sensitivity of 100%-13.2%=86.8%. In practice, it is improbable that all of the 345 FN in node 12 could be corrected with additional information, although new indicators could significantly improve the predictions in other nodes of this tree.

A comparison of model performance would not be complete without accounting for the existing QDiabetes® prognostic model [272]. Figure 6.15 summarises the C-index performance of every model developed in this work, and that of the QDiabetes® models (separately for women and men) evaluated on the same validation cohort from the CPRD data. As expected, the two QDiabetes® models performed, on average, the same as the Cox PH model *without* BG developed in this research. They were marginally outperformed by the NN ensemble model *without* BG, and all four large-data models (Cox PH, ensemble NN, LR, and survival DT) *with* BG.

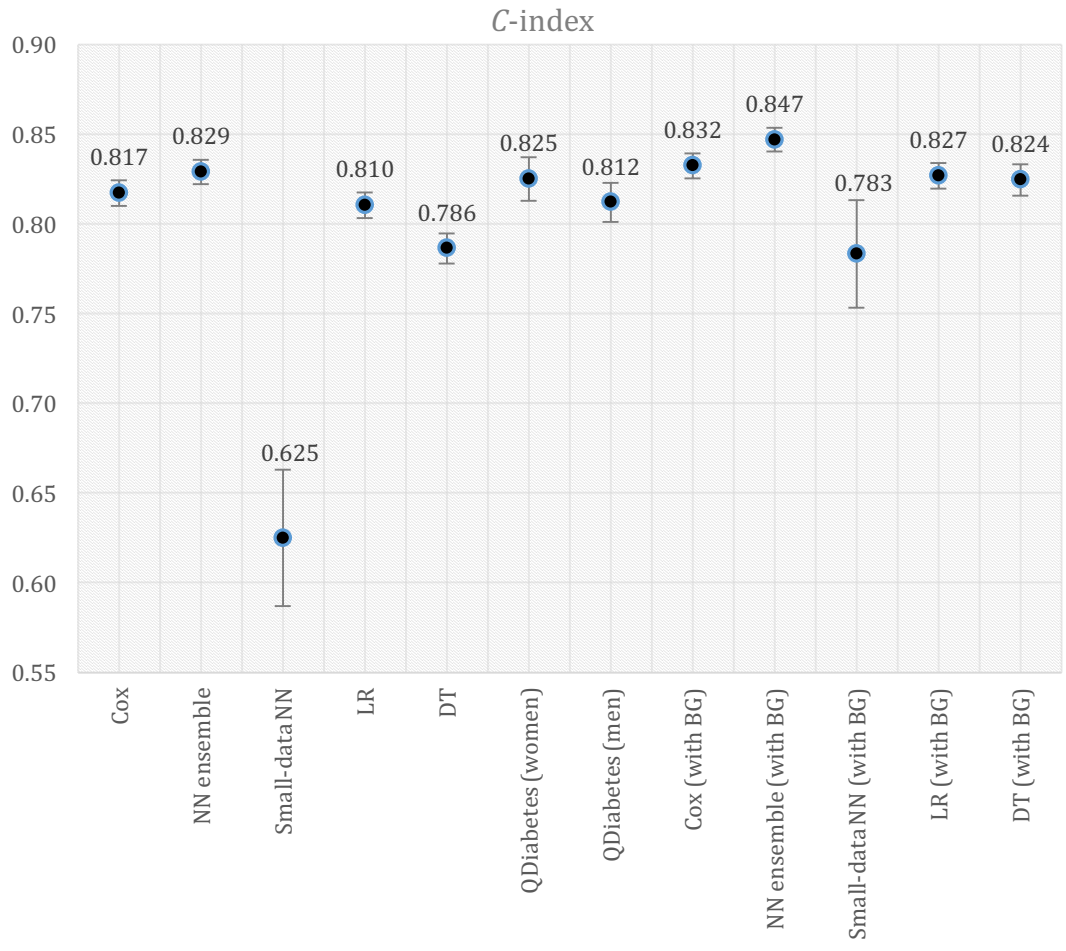


Figure 6.15 Summary of model performance (C-index), including the QDiabetes® model for men and women.

6.5 Chapter conclusions

The key findings demonstrated in this chapter are as follows:

(1) Routinely collected primary care data suffer from complexity, completeness, censoring and consistency challenges ("The 4 Cs"), which limit their potential for data-driven predictive modelling.

(2) The remarkable agreement in concordance and classification accuracies among the Cox PH, LR, NN ensemble, and the existing QDiabestes® model indicate that prognostic performance is a reflection of the quality of the data, rather than a property of an individual model.

(3) Inclusion of available blood glucose data improved all six models. However, this effect was minimal (2% increase in C-index) for the Cox PH, NN ensemble and LR. The improvement was most pronounced for models that did not leverage the imputation of missing variables, i.e. small-data NN (25% increase in C-index) and survival DT (5% increase in C-index). This demonstrates that the potential for using blood glucose data exists, but it remains infeasible until routine blood tests become more frequent in primary care.

(4) In a complete-case scenario of known BG measurements, a small-data NN developed with a balanced subset of 158 samples achieved the same classification accuracy as the Cox PH and NN ensemble models developed on a cohort of 53306 and 19492 samples, respectively. This confirms that good quality data are more important than high performance algorithms or large quantities of incomplete, censored, and imbalanced data.

(5) The survival DT model was able to identify the groups of patients at high 10-year risk of type 2 DM and pin-point those whose data were less conclusive. 82.4% of the DT predictions were concordant. Considering its easily interpretable structure and its ability to handle missing data, the survival DT has the highest practical value among the models explored in this study and is the most appropriate model for the complex task of predicting long-term incidence of a rare disease from routinely collected data.

The task of developing a new generation of dynamic prognostic models for type 2 DM is far from complete. The models prototyped in this work are yet to be evaluated using larger internal and external data. The inherent ability of the NN models to adapt to population dynamics is yet to be quantified with more recent data. Meanwhile, we can rest assured that the existing systems, developed with over 2.5 million records, remain appropriate for the task, until the time when advances in ML for survival modelling and clinical practice for routine data collection builds on the foundation laid by this collaborative work.

Chapter 7

Conclusions

This thesis developed and presented data-efficient ML models for clinical outcome prediction and risk stratification in the context of data limited by size and quality. The three domains considered in this interdisciplinary research were: a) trabecular tissue engineering, b) antibody-incompatible renal transplantation, and c) type 2 diabetes screening in primary care. The (a) trabecular bone and (b) renal transplant datasets, each containing less than 10 observations per predictor variable, exemplified to the extreme the problem of ML modelling from small data. The limited size of available datasets is intrinsic in surgical domains in general, where each sample is a result of costly, invasive, and (fortunately) rare intervention. However, even in domains where large, multi-centre databases are readily available, more data do not necessarily imply proportionally-more information. Routinely collected electronic medical records, such as those involved in (c) diabetes screening in primary care, suffered from limitations in quality that reduced the number of samples available for modelling. The "4 Cs" of routinely collected data, i.e. **c**omplexity, **c**ensoring, in**c**onsistency, and lack of **c**ompleteness, were shown in this thesis to hinder the performance of classical statistical and ML algorithms alike. The methods developed in this thesis enabled the successful deployment of NNs, DTs, and their ensembles despite the limited data and enabled clinical applications that were previously considered beyond the reach of these powerful, yet data-demanding supervised ML algorithms.

7.1 Objectives and the extent to which they were achieved

The objectives of this thesis, as set out in Section 1.6 were met in full. The thesis (1) identified effective strategies for managing data quality limitations in the three abovementioned domains; (2) addressed challenges of learning with limited information by developing and validating a framework for small-data learning (less than 10 observations per predictor variable) with supplementary strategies for incomplete and imbalanced data; (3) designed, implemented, optimised and tested practical NN, DT, and ensemble tools for predictive modelling in the three applications; and (4) used the clinical insights gained from the ML models in order to detect patients at risk and improve short- and long-term individual outcomes.

7.2 Contributions to knowledge

The novel methodological framework for small data developed in this work was motivated by the necessity for scalable predictive models that can make accurate predictions on new observations. The ability of the framework to yield such models was demonstrated for regression NNs by using a large civil engineering study. A small-data NN developed on 41 samples using the proposed framework performed as well as standard NNs trained with a dataset 18 times larger. The remarkable generalising ability (R = 0.87) of the small-data model on 300 new observations confirmed the utility of the proposed framework for producing well-generalising learners despite small data.

The framework comprised: a method of *multiple runs* for model design and optimisation, and a *surrogate data test* for regression model validation in the absence of ample test samples. The method of multiple runs is based on an intuitive principle: rather than solving a complex task with a single learner trained on large amount of data, a *large*

number of learners can be trained with *small amounts of data*, with the anticipation that one of them would excel at the task. By considering performance across a number of learners, the method of multiple runs enabled consistent iterative design optimisation and subsequent model selection. In the absence of ample test data, validating learner performance has proven particularly challenging for regression models, which, unlike classifiers, do not have a pre-defined minimal expected performance threshold. The surrogate data test developed in this thesis addressed this problem by quantifying the lowest expected model performance specific to each dataset. Combined, the two methods enabled the successful application of NNs and DTs for *regression*, *classification*, and *survival modelling* in the three clinical domains.

7.3 Clinical and engineering impact

In trabecular tissue engineering, the framework enabled the development of an accurate NN model for osteoarthritic hip fracture prediction based on an extremely small dataset of 35 trabecular bone samples. This NN estimated, accurate to 0.85 MPa, the trabecular compressive strength in patients suffering from severe hip osteoarthritis by integrating heterogeneous bone scan data with the patient's age and gender. The unique feature of this model was that it was able to achieve 98.3% accurate predictions of bone's mechanical strength from structural and biological parameters without invasive tests. The NN offered a scalable predictive tool for femoral strength in osteoarthritis with a potential extension for other degenerative disorders and to new anatomical locations.

The renal transplant application demonstrated a successful extension of the multiple runs method to small-data DT classification. The method enabled the development of a DT risk stratification model for early transplant rejection based on clinical information from 80 antibody-incompatible recipients. This easy-to-understand, 85% accurate DT

added unprecedented granularity to the Cox PH and LR models, identified key baseline risk factors, and unveiled previously undetermined harmful antibody levels. By integrating known and novel associations, the DT classifier provided a decision support tool for early risk stratification from pre-treatment immunological indicators, leaving clinicians with more time to make essential adjustments to treatment. At the time of publication, this work was the first in the UK to use ML for the prediction of acute rejection from immunological subclass data in antibody-incompatible renal transplantation.

The third and the final application considered in this thesis addressed the problem of *stratifying the 10-year risk of developing type 2 diabetes* in the UK general population from routinely collected primary care data. Several prognostic models, including Cox PH, LR, survival DT, and NN ensembles were successfully developed and validated with 80,000 electronic medical records. Despite consistently achieving 80-85% concordant predictions, the Cox PH, LR, and NN ensembles grossly suffered from the "4 Cs" of routinely collected data, which reduced the amount of available complete samples 100-fold. In comparison, the multiple runs method enabled the development of a 78% concordant NN classifier from only 158 complete-data samples, demonstrating once again that, given adequate data quality, small-data ML techniques can have exceptional prognostic potential in healthcare. The remarkable agreement among the Cox PH, NN ensemble, LR, and the existing QDiabetes® models evidenced that prognostic performance was a reflection of data quality, rather than a property of an individual model.

One sobering implication of this research for the current NHS type 2 diabetes screening system is that substantial improvements to its prognostic value are unlikely to be gained from adopting increasingly more powerful ML algorithms unless more resources can be

dedicated to improving data collection in primary care. Rather than adapting a "store it all" approach, resources should be focused on the quality, completeness and granularity of the observations linked to the outcome of interest. Of a particular practical value is the survival DT model, which – by separating the groups of patients at high and low risk of type 2 diabetes from those patients whose incomplete records required additional granularity – allow for the targeted use of limited NHS resources. As repeatedly evidenced in this research, further improvements to risk stratification and outcome prediction in healthcare do not necessarily depend on large volumes of data. The methodological significance of this research is that it removes the requirement for substantive volumes of data for ML. By lowering the barriers for the application of ML to limited clinical data, this research meaningfully contributes to engineering and clinical practice.

References

[1] A. D. de Groot, *Thought and choice in chess*. The Hague: Mouton Publishers, 1978.

[2] J. Kounios and M. Beeman, "The Cognitive Neuroscience of Insight," *Annu. Rev. Psychol*, vol. 65, no. 4, pp. 71–93, 2014.

[3] D. Somerville, "Spoilers are Criminal," *Brain on Digital*, 2014. [Online]. Available: http://www.brainondigital.com/2014/04/spoilers-are-criminal. [Accessed: 01-Apr-2014].

[4] D. Hardman and L. Macchi, *Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making*. New York: Springer-Verlag, 1986.

[5] M. P. Mattson, "Superior pattern processing is the essence of the evolved human brain," *Front. Neurosci.*, vol. 8, no. 8, 2014.

[6] C. M. Bishop, "Pattern Recognition and Machine Learning," *J. Electron. Imaging*, vol. 16, no. 4, p. 049901, 2007.

[7] E. Alpaydın, *Introduction to machine learning*, 3rd ed. Cambridge: MIT Press, 2014.

[8] T. M. Mitchell, "The Discipline of Machine Learning," *Mach. Learn.*, vol. 17, no. 7, pp. 1–7, 2006.

[9] T. M. Mitchell, *Machine Learning*. London: McGraw-Hill Education, 1997.

[10] T. O. Ayodele, "Types of Machine Learning Algorithms," *New Adv. Mach. Learn.*, pp. 19–49, 2010.

[11] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *J. Big Data*, vol. 3, no. 1, p. 9, 2016.

[12] J. D. Tygar, "Adversarial machine learning," *IEEE Internet Computing*, vol. 15, no. 5. pp. 4–6, 2011.

[13] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. New York: Springer, 2013.

[14] D. L. Hudson and M. E. Cohen, *Neural networks and artificial intelligence for biomedical engineering*. New York: IEEE, 2000.

[15] M. Bagheri, T. N. G. Borhani, A. H. Gandomi, and Z. A. Manan, "A simple modelling approach for prediction of standard state real gas entropy of pure materials.," *SAR QSAR Environ. Res.*, vol. 25, no. 9, pp. 695–710, 2014.

[16] P. Picton, *Neural Networks*, 2nd ed. Basingstoke: Palgrave, 2000.

[17] D. T. Larose and C. D. Larose, *Data Mining and Predictive Analytics*, 2nd ed. Hoboken: Wiley, 2015.

[18] S. Haykin, *Neural Networks and Learning Machines*, 3rd ed. New York: Pearson, 2009.

[19] K. Hornik, M. Stinchcombe, and H. White, "Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks," *Neural Networks*, vol. 3, no. 5, pp. 551–560, 1990.

[20] M. H. Hassoun, *Fundamentals of artificial neural networks*. Cambridge: MIT Press, 1995.

[21] D. Patterson, *Artificial Neural Networks: Theory and Applications*. Singapore: Prentice Hall, 1996.

[22] M. T. Hagan, H. B. Demuth, and M. H. Beale, *Neural network design*. Boston: PWS Pub, 1996.

[23] M. Nielsen, *Neural Networks and Deep Learning*. Determination Press, 2015.

[24] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *2013 IEEE Int. Conf. Acoust. Speech Signal Process.*, no. 6, pp. 6645–6649, 2013.

[25] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[26] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *J. Appl. Biomed.*, vol. 11, pp. 47–58, 2013.

[27] D. M. Vock, J. Wolfson, S. Bandyopadhyay, G. Adomavicius, P. E. Johnson, G. Vazquez-Benitez, and P. J. O'Connor, "Adapting machine learning techniques to censored time-to-event health record data: A general-purpose approach using inverse probability of censoring weighting," *J. Biomed. Inform.*, vol. 61, pp. 119–131, 2016.

[28] M. Zazzi, F. Incardona, M. Rosen-Zvi, M. Prosperi, T. Lengauer, A. Altmann, A. Sonnerborg, T. Lavee, E. Sch??lter, and R. Kaiser, "Predicting response to antiretroviral treatment by machine learning: The euresist project," *Intervirology*, vol. 55, no. 2, pp. 123–127, 2012.

[29] S. Mani, Y. Chen, X. Li, L. Arlinghaus, A. B. Chakravarthy, V. Abramson, S. R. Bhave, M. A. Levy, H. Xu, and T. E. Yankeelov, "Machine learning for predicting the response of breast cancer to neoadjuvant chemotherapy.," *J. Am. Med. Inform. Assoc.*, vol. 20, no. 4, pp. 688–695, 2013.

[30] K. Roe, M. Kakar, T. Seierstad, A. Ree, and D. Olsen, "Early prediction of response to radiotherapy and androgen-deprivation therapy in prostate cancer by repeated functional MRI: a preclinical study," *Radiat. Oncol.*, vol. 6, no. 1, p. 65, 2011.

[31] H. Asadi, R. Dowling, B. Yan, and P. Mitchell, "Machine learning for outcome prediction of acute ischemic stroke post intra-arterial therapy.," *PLoS One*, vol. 9, no. 2, 2014.

[32] H. Saigo, A. Altmann, J. Bogojeska, F. Müller, S. Nowozin, and T. Lengauer, "Learning from past treatments and their outcome improves prediction of in vivo response to anti-HIV therapy.," *Stat. Appl. Genet. Mol. Biol.*, vol. 10, no. 1, 2011.

[33] T. J. Cleophas and A. H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*. New York: Springer, 2015.

[34] I. Inza, B. Calvo, R. Armañanzas, E. Bengoetxea, P. Larrañaga, and J. Lozano, "Machine Learning: An Indispensable Tool in Bioinformatics," in *Bioinformatics Methods in Clinical Research*, vol. 593, R. Matthiesen, Ed. New York: Humana Press, 2010, pp. 25–48.

[35] C. Campbell, "Machine Learning Methodology in Bioinformatics," in *Springer Handbook of Bio-/Neuroinformatics*, N. Kasabov, Ed. Berlin: Springer Berlin Heidelberg, 2014, pp. 185–206.

[36] E. Grossi, "Artificial Neural Networks and Predictive Medicine: a Revolutionary Paradigm Shift," K. Suzuki, Ed. InTech, 2011, pp. 139–150.

[37] K. Hoff, M. Tech, T. Lingner, R. Daniel, B. Morgenstern, and P. Meinicke, "Gene prediction in metagenomic fragments: A large scale machine learning approach," *BMC Bioinformatics*, vol. 9, no. 1, pp. 9–217, 2008.

[38] C. G. DeGroff, S. Bhatikar, J. Hertzberg, R. Shandas, L. Valdes-Cruz, and R. L. Mahajan, "Artificial neural network-based method of screening heart murmurs in children.," *Circulation*, vol. 103, pp. 2711–2716, 2001.

[39] J. Faradmal, A. R. Soltanian, G. Roshanaei, R. Khodabakhshi, and A. Kasaeian, "Comparison of the performance of log-logistic regression and artificial neural networks for predicting breast cancer relapse.," *Asian Pac. J. Cancer Prev.*, vol. 15, no. 14, pp. 5883–5888, 2014.

[40] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nat. Rev. Genet.*, vol. 16, no. 6, pp. 321–332, 2015.

[41] M. K. K. Leung, A. Delong, B. Alipanahi, and B. J. Frey, "Machine learning in genomic medicine: A review of computational problems and data sets," *Proc. IEEE*, vol. 104, no. 1, pp. 176–197, 2016.

[42] N. Wale, "Machine learning in drug discovery and development," *Drug Development Research*, vol. 72, no. 1. pp. 112–119, 2011.

[43] F. Napolitano, Y. Zhao, V. M. Moreira, R. Tagliaferri, J. Kere, M. D'Amato, and D. Greco, "Drug repositioning: A machine-learning approach through data integration," *J. Cheminform.*, vol. 5, no. 6, 2013.

[44] A. Harvey, A. Brand, S. T. Holgate, L. V. Kristiansen, H. Lehrach, A. Palotie, and B. Prainsack, "The future of technologies for personalised medicine," *N. Biotechnol.*, vol. 29, no. 6, pp. 625–633, 2012.

[45] J. A. M. Gray, "The shift to personalised and population medicine," *The Lancet*, vol. 382, no. 9888. pp. 200–201, 2013.

[46] M. de Bruijne, "Machine learning approaches in medical image analysis: From detection to diagnosis," *Medical Image Analysis*, vol. 33, pp. 94–97, 2016.

[47] S. Wang and R. M. Summers, "Machine learning and radiology," *Medical Image Analysis*, vol. 16, no. 5. pp. 933–951, 2012.

[48] M. Vidyasagar, "Machine learning methods in the computational biology of cancer.," *Proc. Math. Phys. Eng. Sci.*, vol. 470, no. 2167, 2014.

[49] X.-B. Yan, W.-Q. Xiong, L. Hu, and K. Zhao, "Cancer prediction based on radical basis function neural network with particle swarm optimization.," *Asian Pac. J. Cancer Prev.*, vol. 15, no. 18, pp. 7775–7780, 2014.

[50] R. Luque-Baena, D. Urda, J. Subirats, L. Franco, and J. Jerez, "Application of genetic algorithms and constructive neural networks for the analysis of microarray cancer data," *Theor. Biol. Med. Model.*, vol. 11, no. Suppl 1, p. S7, 2014.

[51] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. Clifton, and G. D. Clifford, "Machine Learning and Decision Support in Critical Care," *Proc. IEEE*, vol. 104, no. 2, pp. 444–466, 2016.

[52] R. C. Deo, "Machine learning in medicine," *Circulation*, vol. 132, no. 20, pp. 1920–1930, 2015.

[53] C. H. Lee and H.-J. Yoon, "Medical big data: promise and challenges.," *Kidney Res. Clin. Pract.*, vol. 36, no. 1, pp. 3–11, 2017.

[54] O. Badawi, T. Brennan, L. A. Celi, M. Feng, M. Ghassemi, A. Ippolito, A. Johnson, R. G. Mark, L. Mayaud, G. Moody, C. Moses, T. Naumann, M. Pimentel, T. J. Pollard, M. Santos, D. J. Stone, and A. Zimolzak, "Making big data useful for health care: A summary of the inaugural MIT critical data conference," *Journal of Medical Internet Research*, vol. 16, no. 8. p. e22, 2014.

[55] W. Zhu and X. Kan, "Neural Network Cascade Optimizes MicroRNA Biomarker Selection for Nasopharyngeal Cancer Prognosis.," *PLoS One*, vol. 9, no. 10, p. e110537, 2014.

[56] A. Krogh and J. Vedelsby, "Neural Network Ensembles, Cross Validation, and Active Learning," *Adv. Neural Inf. Process. Syst. 7*, pp. 231–238, 1995.

[57] S.-W. Chang, S. Abdul-Kareem, A. Merican, and R. Zain, "Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods," *BMC Bioinformatics*, vol. 14, no. 1, p. 170, 2013.

[58] S. B. Kotsiantis and D. Kanellopoulos, "Data preprocessing for supervised leaning," *Int. J. Comput. Sci.*, vol. 1, no. 2, pp. 111–117, 2006.

[59] A. Xiang, P. Lapuerta, A. Ryutov, J. Buckley, and S. Azen, "Comparison of the performance of neural network methods and Cox regression for censored survival data," *Computational Statistics & Data Analysis*, vol. 34. pp. 243–257, 2000.

[60] J. Buckley and I. James, "Linear regression with censored data," *Biometrika*, vol. 66, no. 3, pp. 429–436, 1979.

[61] D. Faraggi and R. Simon, "Bayesian variable selection method for censored survival data," *Biometrics*, vol. 54, no. 4, pp. 1475–1485, 1998.

[62] Y. Zhou and J. J. McArdle, "Rationale and Applications of Survival Tree and Survival Ensemble Methods," *Psychometrika*, vol. 80, no. 3, pp. 811–833, 2015.

[63] J. Scheffer, "Dealing With Missing Data," *Res. Lett. Inf. Math. Sci.*, vol. 3, pp. 153–160, 2002.

[64] M. L. Brown and J. F. Kros, "Data mining and the impact of missing data," *Ind. Manag. Data Syst.*, vol. 103, no. 8, pp. 611–621, 2003.

[65] R. J. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 2nd ed. Hoboken: John Wiley & Sons, 2002.

[66] Y. Kim, J. Sidney, C. Pinilla, A. Sette, and B. Peters, "Derivation of an amino acid similarity matrix for peptide: MHC binding and its application as a Bayesian prior," *BMC Bioinformatics*, vol. 10, no. 1, p. 394, 2009.

[67] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural Networks*, vol. 21, no. 2–3, pp. 427–436, 2008.

[68] M. M. Rahman and D. N. Davis, "Cluster Based Under-Sampling for Unbalanced Cardiovascular Data," *Proc. World Congr. Eng. 2013*, vol. 3, pp. 1–6, 2013.

[69] A. D. Pozzolo, O. Caelen, and G. Bontempi, "When is undersampling effective in unbalanced classification tasks?," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9284, pp. 200–215.

[70] J. H. Xue and P. Hall, "Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1109–1112, 2015.

[71] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[72] K. J. O'Malley, K. F. Cook, M. D. Price, K. R. Wildes, J. F. Hurdle, and C. M. Ashton, "Measuring diagnoses: ICD code accuracy," *Health Services Research*, vol. 40, no. 5 part 2. pp. 1620–1639, 2005.

[73] J. Langley, S. Stephenson, C. Thorpe, and G. Davie, "Accuracy of injury coding under ICD-9 for New Zealand public hospital discharges.," *Inj. Prev.*, vol. 12, no. 1, pp. 58–61, 2006.

[74] Royal College of Physicians, "A Clinician's Guide to Record Standards – Part 1: Why standardise the structure and content of medical records?" London: Digital and Health Information Policy Directorate, pp. 1–15, 2008.

[75] E. F. Reed, P. Rao, Z. Zhang, H. Gebel, R. A. Bray, I. Guleria, J. Lunz, T. Mohanakumar, P. Nickerson, A. R. Tambur, A. Zeevi, P. S. Heeger, and D. Gjertson, "Comprehensive assessment and standardization of solid phase multiplex-bead arrays for the detection of antibodies to HLA," *Am. J. Transplant.*, vol. 13, no. 7, pp. 1859–1870, 2013.

[76] M. Deisenroth, S. Mohamed, F. Doshi-Velez, A. Krause, and M. Welling, "Data-Efficient Machine Learning," in *ICML 2016 Workshop on Data-Efficient Machine Learning*, 2016, p. 1.

[77] W. Neiswanger and E. Xing, "Efficient Bayesian Inference with Prior Swapping," in *ICML 2016 Workshop on Data-Efficient Machine Learning*, 2016.

[78] R. Antonova, J. Runde, C. Dann, and E. Brunskill, "Improving the Sample Efficiency of Bayesian Optimization Policy Search for Optimal Stopping Problems," in *ICML 2016 Workshop on Data-Efficient Machine Learning*, 2016.

[79] D. Ulmasov, C. Baroukh, B. Chachuat, M. P. Deisenroth, and R. Misener, "Bayesian Optimization with Dimension Scheduling: Application to Biological Systems," *ArXiv,* 1511.05385 [stat.ML], Nov. 2015.

[80] P. Thomas and E. Brunskill, "Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, vol. 48, pp. 2139–2148.

[81] M. Riedmiller, "Neural fitted Q iteration - First experiences with a data efficient neural Reinforcement Learning method," in *Lecture Notes in Computer Science*, 2005, vol. 3720 LNAI, pp. 317–328.

[82] T. Jung and P. Stone, "Gaussian processes for sample efficient reinforcement learning with RMAX-like exploration," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2010, vol. 6321 LNAI, pp. 601–616.

[83] Kaggle, "BigDataCombine," 2013. [Online]. Available: www.kaggle.com/c/battlefin-s-big-data-combine-forecasting-challenge. [Accessed: 10-Apr-2016].

[84] N. A. Khovanova, K. K. Mallick, and T. Shaikhina, "Neural networks for analysis of trabecular bone in osteoarthritis," *Bioinspired, Biomim. Nanobiomaterials*, vol. 4, no. 1, pp. 90–100, 2014.

[85] B. LeBaron and A. S. Weigend, "A bootstrap evaluation of the effect of data splitting on financial time series," *IEEE Trans. Neural Networks*, vol. 9, pp. 213–220, 1998.

[86] G. J. Bowden, "Optimal division of data for neural network models in water resources applications," *Water Resour. Res.*, vol. 38, pp. 1–11, 2002.

[87] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River: Prentice Hall, 1998.

[88] P. D. Wasserman, *Neural computing: theory and practice*. New York: Van Nostrand-Reinhold, 1989.

[89] P. Cunningham, J. Carney, and S. Jacob, "Stability problems with artificial neural networks and the ensemble solution," *Artif. Intell. Med.*, vol. 20, no. 3, pp. 217–225, 2000.

[90] T. A. Holt, N. Khovanova, T. Shaikhina, R. Perera, and A. Fuller, "Identifying risk of diabetes using an artificial neural network applied to primary care data." Project to access the NHS databases, University of Oxford, 2013.

[91] M. Ananda Rao and J. Srinivas, *Neural networks: algorithms and applications*. Pangbourne: Alpha Science International, 2003.

[92] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2323, 1998.

[93] M. E. Harmon and L. C. Baird Iii, "Multi-player residual advantage learning with general function approximation." New Haven: Wright Laboratory, 1996.

[94] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *Science*, vol. 304, no. 78, pp. 78–80, 2004.

[95] J. L. Elman, "Finding structure in time," *Cogn. Sci.*, vol. 14, no. 2, pp. 179–211, 1990.

[96] H. Bourlard and Y. Kamp, "Auto-Association by Multilayer Perceptrons and Singular Value Decomposition," *Biol. Cybern*, vol. 59, pp. 291–294, 1988.

[97] Y. LeCun, M. A. Ranzato, C. Poultney, and S. Chopra, "Efficient Learning of Sparse Representations with an Energy-Based Model," *NIPS*, vol. 1, pp. 1137–1144, 2006.

[98] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 79, no. 8, pp. 2554–2558, 1982.

[99] G. E. Hinton and T. J. Sejnowski, "Learning and relearning in Boltzmann machines," in *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1, Cambridge: MIT Press, 1986, pp. 283–317.

[100] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[101] P. J. Werbos, "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences." Cambridge: Harvard University Press, 1974.

[102] Y. LeCun, "Une procedure d'apprentissage pour reseau a seuil assymetrique (A learning procedure for assymetric threshold networks)," in *Proceedings of Cognitiva 85*, 1985, pp. 599–604.

[103] P. J. Werbos, "Backpropagation Through Time: What It Does and How to Do It," *Proc. IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[104] Y. A. LeCun, L. Bottou, G. B. Orr, and K. R. Müller, "Efficient backpropagation," *Lect. Notes Comput. Sci.*, pp. 9–48, 2012.

[105] S. Dreyfus, "The numerical solution of variational problems," *J. Math. Anal. Appl.*, vol. 5, no. 1, pp. 30–45, 1962.

[106] D. R. Wilson and T. R. Martinez, "The general inefficiency of batch training for gradient descent learning," *Neural Networks*, vol. 16, no. 10, pp. 1429–1451, 2003.

[107] G. Orr and K.-R. Müller, *Neural networks : tricks of the trade*. Berlin, Heidelberg: Springer-Verlag, 1998.

[108] D. Nguyen and B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," *IEEE Int. Jt. Conf. Neural Networks*, vol. 3, p. C21, 1990.

[109] P. J. Werbos, "Generalization of backpropagation with application to a recurrent gas market model," *Neural Networks*, vol. 1, no. 4, pp. 339–356, 1988.

[110] J. J. More, "The Levenberg-Marquardt algorithm: Implementation and theory," *Lect. Notes Math.*, vol. 630, pp. 105–116, 1978.

[111] M. R. Hestenes, *Conjugate Direction Methods in Optimization*. New York, NY: Springer New York, 1980.

[112] H. P. Gavin, "The Levenberg-Marquardt method for nonlinear least squares curve-fitting problems," *Dep. Civ. Environ. Eng. Duke Univ.*, pp. 1–15, 2011.

[113] I. Stich, R. Car, M. Parrinello, and S. Baroni, "Conjugate gradient minimization of the energy functional: A new method for electronic structure calculation," *Phys. Rev. B*, vol. 39, no. 8, 1989.

[114] M. C. Payne, M. P. Teter, D. C. Allan, T. A. Arias, and J. D. Joannopoulos, "Iterative minimization techniques for ab initio total-energy calculations: Molecular dynamics and conjugate gradients," *Rev. Mod. Phys.*, vol. 64, no. 4, pp. 1045–1097, 1992.

[115] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.

[116] J. Bruck and J. W. Goodman, "A Generalized Convergence Theorem for Neural Networks," *IEEE Trans. Inf. Theory*, vol. 34, no. 5, pp. 1089–1092, 1988.

[117] W. Wu, G. Feng, Z. Li, and Y. Xu, "Deterministic convergence of an online gradient method for BP neural networks," *IEEE Trans. Neural Networks*, vol. 16, no. 3, pp. 533–540, 2005.

[118] S. Lawrence, C. L. Giles, and A. C. Tsoi, "What Size Neural Network Gives Optimal Generalization ? Convergence Properties of Backpropagation." Technical Reports from UMIACS, University of Maryland, pp. 1–37, 1998.

[119] P. Baldi and P. Sadowski, "The Dropout Learning Algorithm.," *Artif. Intell.*, vol. 210, pp. 78–122, 2014.

[120] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. New York: Chapman & Hall, 1984.

[121] V. Podgorelec, P. Kokol, B. Stiglic, and I. Rozman, "Decision trees: An overview and their use in medicine," *J. Med. Syst.*, vol. 26, no. 5, pp. 445–463, 2002.

[122] M. D. Murphy, M. J. O'Mahony, L. Shalloo, P. French, and J. Upton, "Comparison of modelling techniques for milk-production forecasting.," *J. Dairy Sci.*, vol. 97, no. 6, pp. 3352–3363, 2014.

[123] R. J. Marshall, "The use of classification and regression trees in clinical epidemiology," *J. Clin. Epidemiol.*, vol. 54, no. 6, pp. 603–609, 2001.

[124] A. Azar and S. El-Metwally, "Decision tree classifiers for automated medical diagnosis," *Neural Comput. Appl.*, vol. 23, no. 7–8, pp. 2387–2403, 2013.

[125] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision Forests for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning," *Learning*, vol. 7, pp. 81–227, 2011.

[126] L. Ceriani and P. Verme, "The origins of the Gini index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini," *J. Econ. Inequal.*, vol. 10, no. 3, pp. 421–443, 2012.

[127] L. E. Raileanu and K. Stoffel, "Theoretical Comparison between the Gini Index and Information Gain Criteria," *Ann. Math. Artif. Intell.*, vol. 41, no. 1, pp. 77–93, 2004.

[128] A. Dobra and J. Gehrke, "Bias Correction in Classification Tree Construction," in *Proceedings of the 18th International Conference on Machine Learning*, 2001, pp. 90–97.

[129] V. Voinov, M. Nikulin, and N. Balakrishnan, *Chi-Squared Goodness of Fit Tests with Applications*. Waltham: Academic Press, 2013.

[130] A. A. Freitas, "A survey of evolutionary algorithms for data mining and knowledge discovery," in *Advances in Evolutionary Computation*, Berlin: Springer Berlin Heidelberg, 2003, pp. 819–845.

[131] J. R. Quinlan, *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kaufmann Publishers, 1993.

[132] D. Opitz and R. Maclin, "Popular Ensemble Methods: An Empirical Study," *J. Artif. Intell. Res.*, vol. 11, pp. 169–198, 1999.

[133] T. G. Dietterich, "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, 2000.

[134] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[135] B. Parmanto, P. W. Munro, and H. R. Doyle, "Reducing Variance of Committee Prediction with Resampling Techniques," *Conn. Sci.*, vol. 8, no. 3–4, pp. 405–425, 1996.

[136] Y. Freund and R. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," *Comput. Learn. theory*, vol. 55, pp. 119–139, 1995.

[137] Z. Ahmad and J. Zhang, "A comparison of different methods for combining multiple neural networks models," *Proceedings of the 2002 International Joint Conference on Neural Networks IJCNN02 Cat No02CH37290*, vol. 1. pp. 828–833, 2002.

[138] L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, no. 5, pp. 1–35, 1999.

[139] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. New York: Springer, 2009.

[140] S. C. Bagley, H. White, and B. A. Golomb, "Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain," *J. Clin. Epidemiol.*, vol. 54, no. 10, pp. 979–985, 2001.

[141] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*, 3rd ed., no. Wiley series in probability and statistics. Hoboken: Wiley, 2013.

[142] S. F. Chan, J. J. Deeks, P. Macaskill, and L. Irwig, "Three methods to construct predictive models using logistic regression and likelihood ratios to facilitate adjustment for pretest probability give similar results," *J. Clin. Epidemiol.*, vol. 61, no. 1, pp. 52–63, 2008.

[143] D. L. Simel, G. P. Samsa, and D. B. Matchar, "Likelihood ratios with confidence: Sample size estimation for diagnostic test studies," *J. Clin. Epidemiol.*, vol. 44, no. 8, pp. 763–770, 1991.

[144] G. King and L. Zeng, "Logistic Regression in Rare Events Data," *Polit. Anal.*, vol. 9, no. 2, pp. 137–163, 2001.

[145] D. M. Potter, "A permutation test for inference in logistic regression with small- and moderate-sized data sets," *Stat. Med.*, vol. 24, no. 5, pp. 693–708, 2005.

[146] J. M. Bland, "An Introduction to Medical Statistics, 3rd edition," New York: Oxford Univ. Press, pp. 137–155, 2000.

[147] D. R. Cox, "Regression models and life tables," *J. R. Stat. Soc. Ser. B*, vol. 34, no. 2, pp. 187–220, 1972.

[148] R. Miller and J. Halpern, "Regression with censored data," *Biometrika*, vol. 69, pp. 521–531, 1982.

[149] J. D. Singer and J. B. Willett, "Fitting Cox Regression Models," in *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*, Oxford: Oxford University Press, 2009, pp. 1–644.

[150] J. Cohen, *Applied Multiple Regression: Correlation Analysis for the Behavioral Sciences*, 3rd ed. London: L. Erlbaum Associates, 2003.

[151] R. Routledge, "Fisher's Exact Test," *Encycl. Biostat.*, vol. 53, pp. 1961–1964, 2005.

[152] C. J. Wild and G. A. Seber, "The Wilcoxon Rank-Sum Test," in *Chance Encounters: A First Course in Data Analysis and Inference*, New York: John Wiley & Sons, 2000, p. 611.

[153] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Stat. Comput.*, vol. 21, pp. 137–146, 2009.

[154] M. S. Schroeder, J. Culhane, A. C. Quackenbush, and B. Haibe-Kains, "survcomp: an R/Bioconductor package for performance assessment and comparison of survival models," *Bioinformatics*, vol. 27, no. 22. pp. 3206–3208, 2011.

[155] T. A. Gerds, "Product-Limit Estimation for Censored Event History Analysis," 2017. [Online]. Available: https://cran.r-project.org/package=prodlim. [Accessed: 11-Mar-2017].

[156] S. R. Lele, J. L. Keim, and P. Solymos, "ResourceSelection: Resource Selection (Probability) Functions for Use-Availability Data." 2017. [Online]. Available : https://cran.r-project.org/package=ResourceSelection. [Accessed 11-Mar-2017].

[157] M. Kuhn, "CARET: Classification and Regression Training." Astrophysics Source Code Library, 1:05003, 2015.

[158] L. Torgo, "Data Mining with R, learning with case studies." New York: Chapman and Hall, 2010.

[159] T. M. Therneau, "A Package for Survival Analysis in S." 2015. [Online]. Available: https://cran.r-project.org/package=survival. [Accessed: 11-Mar-2017].

[160] J. A. Huwaldt, "Plot Digitizer." 2012. [Online]. Available: http://plotdigitizer. sourceforge.net. [Accessed: 1-Jan-2013].

[161] P. H. Sherrod, "DTREG." 2014. [Online]. Available: www.dtreg.com [Accessed: 15-Nov-2013].

[162] The Mendeley Support Team, Mendeley Ltd, P. Dearden, B. Kowalski, J. Lowe, R. Roland, M. Surridge, S. Thomas, and S. Jones, "Mendeley Desktop," *Mendeley Desktop*. pp. 1–16, 2011.

[163] "Performing Asynchronous Speech Recognition," *Google Cloud Speech API Documentation*, 2017. [Online]. Available: https://cloud.google.com/speech/docs /async-recognize. [Accessed: 28-Aug-2017].

[164] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cem. Concr. Res.*, vol. 28, no. 12, pp. 1797–1808, 1998.

[165] I.-C. Yeh, "UCI Machine Learning Repository: Concrete Compressive Strength Data Set," *Machine Learning Repository, University of California Irvine, Center of Machine Learning and Intelligent Systems*, 2007. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength.

[166] E. Perilli, M. Baleani, C. Öhman, F. Baruffaldi, M. Viceconti, C. Ohman, F. Baruffaldi, and M. Viceconti, "Structural parameters and mechanical strength of cancellous bone in the femoral head in osteoarthritis do not depend on age," *Bone*, vol. 41, no. 5, pp. 760–768, 2007.

[167] N. Khovanova, S. Daga, T. Shaikhina, N. Krishnan, J. Jones, D. Zehnder, D. Mitchell, R. Higgins, D. Briggs, and D. Lowe, "Subclass analysis of donor HLA-specific IgG in antibody-incompatible renal transplantation reveals a significant association of IgG4 with rejection and graft failure," *Transpl. Int.*, vol. 28, no. 12, pp. 1405–1415, 2015.

[168] D. Lowe, R. Higgins, D. Zehnder, and D. C. Briggs, "Significant IgG subclass heterogeneity in HLA-specific antibodies: Implications for pathogenicity, prognosis, and the rejection response," *Hum. Immunol.*, vol. 74, no. 5, pp. 666–672, 2013.

[169] G. Forman and I. Cohen, "Learning from Little: Comparison of Classifiers Given Little Training," *Proc PKDD*, vol. 19, pp. 161–172, 2004.

[170] D.-C. Li, C.-S. Wu, T.-I. Tsai, and Y.-S. Lina, "Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge," *Comput. Oper. Res.*, vol. 34, no. 4, pp. 966–982, 2007.

[171] R. Lanouette, J. Thibault, and J. L. Valade, "Process modeling with neural networks using small experimental datasets," *Comput. Chem. Eng.*, vol. 23, no. 9, pp. 1167–1176, Nov. 1999.

[172] Z. Ghahramani and M. Jordan, "Learning from incomplete data," *Lab Memo No. 1509, CBCL Paper No. 108, MIT AI Lab*. 1994.

[173] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[174] A. Gelman and J. Hill, "Missing-data imputation," in *Data Analysis Using Regression and Multilevel/Hierarchical Models, Analytical Methods for Social Research*, Cambridge University Press, 2006, pp. 529–544.

[175] J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls.," *BMJ*, vol. 338, p. b2393, 2009.

[176] S. Van Buuren, H. C. Boshuizen, and D. L. Knook, "Multiple imputation of missing blood pressure covariates in survival analysis," *Stat. Med.*, vol. 18, no. 6, pp. 681–694, 1999.

[177] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: a critical evaluation," *BMC Med. Inform. Decis. Mak.*, vol. 16, no. S3, p. 74, 2016.

[178] S. Van Buuren and K. Groothuis-Oudshoorn, "Multivariate Imputation by Chained Equations," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.

[179] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: Issues and guidance for practice," *Stat. Med.*, vol. 30, no. 4, pp. 377–399, 2011.

[180] J. Barnard and X.-L. Meng, "Applications of multiple imputation in medical studies: from AIDS to NHANES," *Stat. Methods Med. Res.*, vol. 8, no. 1, pp. 17–36, 1999.

[181] T. M. Therneau and E. J. Atkinson, "An Introduction to Recursive Partitioning using the rpart Routine," *Stats*, vol. 116, no. 61. pp. 1–52, 2017.

[182] C. Ferri, J. Hernández-Orallo, and R. Modroiu, "An experimental comparison of performance measures for classification," *Pattern Recognit. Lett.*, vol. 30, no. 1, pp. 27–38, 2009.

[183] L. a. Jeni, J. F. Cohn, and F. De La Torre, "Facing Imbalanced Data--Recommendations for the Use of Performance Metrics," *2013 Hum. Assoc. Conf. Affect. Comput. Intell. Interact.*, pp. 245–251, 2013.

[184] J. Pastor-Pellicer, F. Zamora-Martínez, S. España-Boquera, and M. J. Castro-Bleda, "F-measure as the error function to train neural networks," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2013, vol. 7902 LNCS, no. 1, pp. 376–384.

[185] X. Lu, K. Tang, and X. Yao, "Evolving Neural Networks with Maximum AUC for Imbalanced Data Classification," *Hybrid Artif. Intell. Syst.*, vol. 6076, pp. 335–342, 2010.

[186] Y. Nan, K. M. Chai, W. S. Lee, and H. L. Chieu, "Optimizing F-measure: A Tale of Two Approaches," *Proc. 29th Int. Conf. Mach. Learn.*, pp. 289–296, 2012.

[187] M. Jansche, "Maximum expected F-measure training of logistic regression models," *Proc. 2005 Conf. Hum. Lang. Technol. Empir. Methods Nat. Lang. Process.*, no. October, pp. 692–699, 2005.

[188] C. Drummond and R. C. Holte, "C4.5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling," *Proc. Int. Conf. Mach. Learn.*, pp. 1–8, 2003.

[189] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[190] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "MUTE: Majority under-sampling technique," in *ICICS 2011 - 8th International Conference on Information, Communications and Signal Processing*, 2011.

[191] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proceedings of the International Joint Conference on Neural Networks*, 2008, pp. 1322–1328.

[192] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory Undersampling for Class Imbalance Learning," *IEEE Trans. Syst. Man Cybern.*, vol. 39, no. 2, pp. 539–550, 2009.

[193] T. Schreiber and A. Schmitz, "Improved Surrogate Data for Nonlinearity Tests," *Phys. Rev. Lett.*, vol. 77, no. 4, pp. 635–638, 1996.

[194] J. Timmer, "Power of surrogate data testing with respect to nonstationarity," *Phys. Rev. E*, vol. 58, no. 4, pp. 5153–5156, 1998.

[195] J. L. Johnson, *Probability and Statistics for Computer Science*. New York: Wiley, 2011.

[196] K. Levenberg, "A Method for the Solution of Certain Non-linear Problems in Least-Squares," *Q. Appl. Math.*, vol. 2, pp. 164–168, 1944.

[197] D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *J. Soc. Ind. Appl. Math.*, vol. 11, pp. 431–441, 1963.

[198] A. Pasini, "Artificial neural networks for small dataset analysis," *J. Thorac. Dis.*, vol. 7, no. 5, pp. 953–960, 2015.

[199] C. Shu and D. H. Burn, "Artificial neural network ensembles and their application in pooled flood frequency analysis," *Water Resources Research*, vol. 40, no. 9. 2004.

[200] D. J. C. MacKay, "Bayesian Methods for Adaptive Models," Calif. Inst. of Technol., Pasadena, 1992.

[201] M. Y. Chan, J. R. Center, J. A. Eisman, and T. V. Nguyen, "Bone mineral density and association of osteoarthritis with fracture risk," *Osteoarthritis Cartilage*, vol. 22, no. 9, pp. 1251–1258, 2014.

[202] M. Bessho, I. Ohnishi, H. Okazaki, W. Sato, H. Kominami, S. Matsunaga, and K. Nakamura, "Prediction of the strength and fracture location of the femoral neck by CT-based finite-element method: A preliminary study on patients with hip fracture," *J. Orthop. Sci.*, vol. 9, pp. 545–550, 2004.

[203] J. H. Keyak, S. A. Rossi, K. A. Jones, and H. B. Skinner, "Prediction of femoral fracture load using automated finite element modeling," *J. Biomech.*, vol. 31, pp. 125–133, 1997.

[204] B. Helgason, E. Perilli, E. Schileo, F. Taddei, S. Brynjolfsson, and M. Viceconti, "Mathematical relationships between bone density and mechanical properties: A literature review," *Clin. Biomech.*, vol. 23, pp. 135–146, 2008.

[205] O. Johnel, B. Gullberg, E. Allander, and J. A. Kanis, "The apparent incidence of hip fracture in Europe: A study of national register sources," *Osteoporos. Int.*, vol. 2, no. 6, pp. 298–302, 1992.

[206] D. K. Dhanwal, E. M. Dennison, N. C. Harvey, and C. Cooper, "Epidemiology of hip fracture: Worldwide geographic variation," *Indian J. Orthop.*, vol. 45, no. 1, pp. 15–22, 2011.

[207] E. Guerado, R. M. Sandalio, Z. Caracuel, and E. Caso, "Understanding the pathogenesis of hip fracture in the elderly, osteoporotic theory is not reflected in the outcome of prevention programmes," *World J. Orthop.*, vol. 7, no. 4, p. 218, 2016.

[208] R. Bartl and B. Frisch, "Pathogenesis of Osteoporosis," in *Osteoporosis*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 29–37.

[209] K. Sinusas, "Osteoarthritis: Diagnosis and Treatment," *Am. Fam. Physician*, vol. 1, no. 86, pp. 49–56, 2012.

[210] Arthritis Research UK, "Prevalence of osteoarthritis in England and local authorities: Warwickshire," *Public Health England*. p. 4, 2014.

[211] A. Stewart and A. J. Black, "Bone mineral density in osteoarthritis.," *Curr. Opin. Rheumatol.*, vol. 12, no. 5, pp. 464–467, 2000.

[212] V. V. Živković, B. B. Stamenković, J. Nedović, A. Dimić, and K. Marković, "Bone Mineral Density in Osteoarthritis," *Sci. J. Fac. Med. Niš*, vol. 27, no. 3, pp. 135–141, 2010.

[213] G. J. Crane, N. L. Fazzalari, I. H. Parkinson, and B. Vernon-Roberts, "Age-related changes in femoral trabecular bone in arthrosis.," *Acta Orthop. Scand.*, vol. 61, no. 5, pp. 421–426, 1990.

[214] M. V. Foss and P. D. Byers, "Bone density, osteoarthrosis of the hip, and fracture of the upper end of the femur," *Ann. Rheum. Dis.*, vol. 31, pp. 259–264, 1972.

[215] N. K. Arden, S. Crozier, H. Smith, F. Anderson, C. Edwards, H. Raphael, and C. Cooper, "Knee pain, knee osteoarthritis, and the risk of fracture.," *Arthritis Rheum.*, vol. 55, no. 4, pp. 610–615, 2006.

[216] N. C. Wright, J. R. Lisse, B. T. Walitt, et al., "Arthritis increases the risk for fractures - Results from the women's health initiative," *J. Rheumatol.*, vol. 38, no. 8, pp. 1680–1688, 2011.

[217] R. Bartl, B. Frisch, and C. Bartl, "Biology of Bone," in *Osteoporosis: Diagnosis, Prevention, Therapy*, 2nd ed., Berlin: Springer-Verlag, 2009, pp. 7–28.

[218] T. Hildebrand, A. Laib, R. Müller, J. Dequeker, and P. Rüegsegger, "Direct Three-Dimensional Morphometric Analysis of Human Cancellous Bone: Microstructural Data from Spine, Femur, Iliac Crest, and Calcaneus," *J. Bone Miner. Res.*, vol. 14, no. 7, pp. 1167–1174, 1999.

[219] L. J. Gibson, M. F. Ashby, and B. A. Harley, *Cellular materials in nature and medicine*. Cambridge, MA: Cambridge University Press, 2010.

[220] L. J. Gibson, "The mechanical behaviour of cancellous bone," *J. Biomech.*, vol. 18, no. 5, pp. 317–328, 1985.

[221] M. Stauber and R. Müller, "Age-related changes in trabecular bone microstructures: Global and local morphometry," *Osteoporos. Int.*, vol. 17, no. 4, pp. 616–626, 2006.

[222] J. Galante, W. Rostoker, and R. D. Ray, "Physical properties of trabecular bone," *Calcif. Tissue Res.*, vol. 5, no. 1, pp. 236–246, 1970.

[223] E. Perilli, M. Baleani, C. Öhman, R. Fognani, F. Baruffaldi, and M. Viceconti, "Dependence of mechanical compressive strength on local variations in microarchitecture in cancellous bone of proximal human femur," *J. Biomech.*, vol. 41, no. 2, pp. 438–446, 2008.

[224] J. Kabel, B. Van Rietbergen, A. Odgaard, and R. Huiskes, "Constitutive relationships of fabric, density, and elastic properties in cancellous bone architecture," *Bone*, vol. 25, no. 4, pp. 481–486, 1999.

[225] M. J. Silva, T. M. Keaveny, and W. C. Hayes, "Computed tomography-based finite element analysis predicts failure loads and fracture patterns for vertebral sections," *J. Orthop. Res.*, vol. 16, no. 3, pp. 300–308, 1998.

[226] K. K. Mallick, J. Winnett, W. van Grunsven, J. Lapworth, and G. C. Reilly, "Three-dimensional porous bioscaffolds for bone tissue regeneration: Fabrication via adaptive foam reticulation and freeze casting techniques, characterization, and cell study," *J. Biomed. Mater. Res. Part A*, vol. 100A, no. 11, pp. 2948–2959, 2012.

[227] S. Bose, M. Roy, and A. Bandyopadhyay, "Recent advances in bone tissue engineering scaffolds," *Trends Biotechnol.*, vol. 30, no. 10, pp. 546–554, 2012.

[228] H. Yonaba, F. Anctil, and V. Fortin, "Comparing Sigmoid Transfer Functions for Neural Network," *J. Hydrol. Eng.*, vol. 15, no. 4, pp. 275–283, 2010.

[229] A. A. Zadpoor, G. Campoli, and H. Weinans, "Neural network prediction of load from the morphology of trabecular bone," *Appl. Math. Model.*, vol. 37, no. 7, pp. 5260–5276, 2013.

[230] R. Hambli, "Apparent damage accumulation in cancellous bone using neural networks," *J. Mech. Behav. Biomed. Mater.*, vol. 4, no. 6, pp. 868–878, 2011.

[231] T. Shaikhina, N. A. Khovanova, and K. K. Mallick, "Artificial neural networks in hard tissue engineering: another look at age-dependence of trabecular bone properties in osteoarthritis," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 2014, pp. 622–625.

[232] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Machine Learning for Predictive Modelling Based on Small Data in Biomedical Engineering," in *IFAC-PapersOnLine*, 2015, vol. 48, no. 20, pp. 469–474.

[233] T. Shaikhina and N. A. Khovanova, "Handling limited datasets with neural networks in medical applications: a small-data approach," *Artif. Intell. Med.*, vol. 75, pp. 51–63, 2017.

[234] L. Geris, *Computational Modeling in Tissue Engineering*. Berlin: Springer, 2013.

[235] C. Eller-Vainicher, V. V Zhukouskaya, Y. V Tolkachev, S. S. Koritko, E. Cairoli, E. Grossi, P. Beck-Peccoz, I. Chiodini, and A. P. Shepelkevich, "Low bone mineral density and its predictors in type 1 diabetic patients evaluated by the classic statistics and artificial neural network analysis.," *Diabetes Care*, vol. 34, no. 10, pp. 2186–20191, 2011.

[236] D. Peteiro-Barral, V. Bolon-Canedo, A. Alonso-Betanzos, B. Guijarro-Berdinas, and N. Sanchez-Marono, "Toward the scalability of neural networks through feature selection," *Expert Syst. Appl.*, vol. 40, pp. 2807–2816, 2013.

[237] NHS Blood and Transplant, "Organ Donation and Transplantation: Activity Report 2016/17," *National Health Service*. p. 167, 2017. [Online]. Available: https://www.odt.nhs.uk/statistics-and-reports/annual-activity-report/. [Accessed: 13-Jul-2017].

[238] R. A. Montgomery, B. E. Lonze, K. E. King, E. S. Kraus, L. M. Kucirka, J. E. Locke, D. S. Warren, C. E. Simpkins, N. N. Dagher, A. L. Singer, A. A. Zachary, and D. L. Segev, "Re: Desensitization in HLA-incompatible kidney recipients and survival," *Journal of Urology*, vol. 187, no. 5. pp. 1766–1767, 2012.

[239] R. Higgins, D. Zehnder, K. Chen, D. Lowe, J. McKinnell, F. T. Lam, H. Kashi, L. C. Tan, C. Imray, S. Fletcher, N. Krishnan, R. Hamer, and D. Briggs, "The histological development of acute antibody-mediated rejection in HLA antibody-incompatible renal transplantation.," *Nephrol. Dial. Transplant*, vol. 25, no. 4, pp. 1306–1312, 2010.

[240] T. S. Purnell, P. Auguste, D. C. Crews, J. Lamprea-Montealegre, T. Olufade, R. Greer, P. Ephraim, J. Sheu, D. Kostecki, N. R. Powe, H. Rabb, B. Jaar, and L. E. Boulware, "Comparison of life participation activities among adults treated by hemodialysis, peritoneal dialysis, and kidney transplantation: a systematic review.," *Am. J. Kidney Dis.*, vol. 62, no. 5, pp. 953–973, 2013.

[241] R. M. Higgins, D. J. Bevan, B. S. Carey, C. K. Lea, M. Fallon, R. Bühler, R. W. Vaughan, P. J. O'Donnell, S. A. Snowden, M. Bewick, and B. M. Hendry, "Prevention of hyperacute rejection by removal of antibodies to HLA immediately before renal transplantation," *Lancet*, vol. 348, no. 9036, pp. 1208–1211, 1996.

[242] R. M. Higgins, D. J. Bevan, R. W. Vaughan, A. O. Phillips, S. Snowden, M. Bewick, J. E. Scoble, and B. M. Hendry, "5-Year follow-up of patients successfully transplanted after immunoadsorption to remove anti-HLA antibodies," *Nephron*, vol. 74, no. 1, pp. 53–57, 1996.

[243] R. Higgins, D. Lowe, M. Hathaway, F. Lam, H. Kashi, L. C. Tan, C. Imray, S. Fletcher, K. Chen, N. Krishnan, R. Hamer, D. Zehnder, and D. Briggs, "Rises and falls in donor-specific and third-party HLA antibody levels after antibody incompatible transplantation.," *Transplantation*, vol. 87, no. 6, pp. 882–888, 2009.

[244] R. Higgins, D. Lowe, M. Hathaway, F. T. Lam, H. Kashi, L. C. Tan, C. Imray, S. Fletcher, K. Chen, N. Krishnan, R. Hamer, D. Zehnder, and D. Briggs, "Double filtration plasmapheresis in antibody-incompatible kidney transplantation," *Ther. Apher. Dial.*, vol. 14, no. 4, pp. 392–399, 2010.

[245] R. a. Montgomery, D. S. Warren, D. L. Segev, and A. a. Zachary, "HLA incompatible renal transplantation," *Curr. Opin. Organ Transplant.*, vol. 17, no. 4, pp. 386–392, 2012.

[246] T. J. Kindt, R. A. Goldsby, B. A. Osborne, and J. Kuby, *Kuby Immunology*, 6th ed. New York: W. H. Freeman, 2007.

[247] G. Hönger, H. Hopfer, M.-L. Arnold, B. M. Spriewald, S. Schaub, and P. Amico, "Pretransplant IgG subclasses of donor-specific human leukocyte antigen antibodies and development of antibody-mediated rejection.," *Transplantation*, vol. 92, no. 1, pp. 41–47, 2011.

[248] M. L. Arnold, I. S. Ntokou, I. I. N. Doxiadis, B. M. Spriewald, J. N. Boletis, and A. G. Iniotaki, "Donor-specific HLA antibodies: Evaluating the risk for graft loss in renal transplant recipients with isotype switch from complement fixing IgG1/IgG3 to noncomplement fixing IgG2/IgG4 anti-HLA alloantibodies," *Transpl. Int.*, vol. 27, no. 3, pp. 253–261, 2014.

[249] P. Stastny, S. Ring, C. Lu, J. Arenas, M. Han, and B. Lavingia, "Role of immunoglobulin (Ig)-G and IgM antibodies against donor human leukocyte antigens in organ transplant recipients," *Hum. Immunol.*, vol. 70, no. 8, pp. 600–604, 2009.

[250] E. J. Griffiths, R. E. Nelson, P. J. Dupont, and A. N. Warrens, "Skewing of pretransplant anti-HLA class I antibodies of immunoglobulin G isotype solely toward immunoglobulin G1 subclass is associated with poorer renal allograft survival.," *Transplantation*, vol. 77, no. 11, pp. 1771–1783, 2004.

[251] R. Greco, T. Papalia, D. Lofaro, S. Maestripieri, D. Mancuso, and R. Bonofiglio, "Decisional Trees in Renal Transplant Follow-up," *Transplant. Proc.*, vol. 42, no. 4, pp. 1134–1136, 2010.

[252] S. Krikov, A. Khan, B. C. Baird, L. L. Barenbaum, A. Leviatov, J. K. Koford, and A. S. Goldfarb-Rumyantzev, "Predicting kidney transplant survival using tree-based modeling," *J. Am. Soc. Artif. Intern. Organs*, vol. 53, no. 5, pp. 592–600, 2007.

[253] A. Decruyenaere, P. Decruyenaere, P. Peeters, F. Vermassen, T. Dhaene, and I. Couckuyt, "Prediction of delayed graft function after kidney transplantation: comparison between logistic regression and machine learning methods," *BMC Med. Inform. Decis. Mak.*, vol. 15, no. 83, 2015.

[254] D. Lofaro, S. Maestripieri, R. Greco, T. Papalia, D. Mancuso, D. Conforti, and R. Bonofiglio, "Prediction of chronic allograft nephropathy using classification trees," *Transplant. Proc.*, vol. 42, no. 4, pp. 1130–1133, 2010.

[255] N. A. Khovanova, D. P. Lowe, S. Daga, T. Shaikhina, D. A. Mitchell, D. Zehnder, D. Briggs, and R. Higgins, "Assessment of IgG subclass significance for early graft rejection and long-term survival in HLA-antibody incompatible renal transplantation: multivariate approach," *Am. J. Transplant.*, vol. 15, no. S3, p. 338, 2015.

[256] T. Shaikhina, D. Lowe, S. Daga, D. Briggs, R. Higgins, and N. Khovanova, "Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation," *Biomedical Signal Processing and Control*. 2017.

[257] T. Shaikhina, N. Khovanova, S. Daga, N. Krishnan, D. P. Lowe, D. A. Mitchell, D. Briggs, and R. Higgins, "Prediction of acute antibody mediated rejection in antibody incompatible renal transplantation using machine learning for wide data," *Am. J. Transplant.*, vol. 15, no. S3, p. 1364, 2015.

[258] N. Goudarzi, D. Shahsavani, F. Emadi-Gandaghi, and M. A. Chamjangali, "Application of random forests method to predict the retention indices of some polycyclic aromatic hydrocarbons.," *J. Chromatogr. A*, vol. 1333, pp. 25–31, Mar. 2014.

[259] G. Dieplinger, V. Ditt, W. Arns, A. Huppertz, T. Kisner, M. Hellmich, U. Bauerfeind, and D. L. Stippel, "Impact of de novo donor-specific HLA antibodies detected by Luminex solid-phase assay after transplantation in a group of 88 consecutive living-donor renal transplantations," *Transpl. Int.*, vol. 27, no. 1, pp. 60–68, 2014.

[260] C. Lefaucheur, A. Loupy, G. S. Hill, J. Andrade, D. Nochy, C. Antoine, C. Gautreau, D. Charron, D. Glotz, and C. Suberbielle-Boissel, "Preexisting donor-specific HLA antibodies predict outcome in kidney transplantation.," *J. Am. Soc. Nephrol.*, vol. 21, no. 8, pp. 1398–1406, 2010.

[261] R. Higgins, N. Khovanova, and N. Krishnan, "UK Transplant Registry Application for Data." p. 4, 2015.

[262] World Health Organization, "Global Report on Diabetes," *WHO Library Cataloguing-in-Publication Data*. pp. 1–88, 2016. [Online]. Available: http://www.who.int . [Accessed: 10-Jan-2017].

[263] R. Bilous and R. Donnelly, *Handbook of Diabetes*, 4th ed. Oxford: Wiley-Blackwell, 2010.

[264] Health and Social Care Information Centre, "Quality and Outcomes Framework – Prevalence, Achievements and Exceptions Report: England, 2015-16," *NHS Digital*. p. 26, 2016.

[265] Diabetes UK, "Diabetes: Facts and Stats." pp. 1–17, 2016. [Online]. www.diabetes.org.uk/professionals/position-statements-reports/statistics. [Accessed: 13-Jan-2017].

[266] M. B. Weber and K. M. V. Narayan, "Preventing type 2 diabetes: Genes or lifestyle?," *Primary Care Diabetes*, vol. 2, no. 2, pp. 65–66, 2008.

[267] National Institute for Health and Care Excellence, "Type 2 diabetes : prevention in people at high risk," Manchester: NICE guidelines, pp. 1–154, 2012.

[268] National Institute for Health and Care Excellence, "Type 2 diabetes in adults: management," Manchester: NICE guidelines, pp. 1–44, 2015.

[269] T. A. Chowdhury and R. Shah, "Reducing the risk of diabetes:," *BMJ*, vol. 351, no. h4595, 2015.

[270] R. Yan, X. Wang, L. Huang, J. Lin, W. Cai, and Z. Zhang, "GPCRserver: an accurate and novel G protein-coupled receptor predictor.," *Mol. Biosyst.*, vol. 10, no. 10, pp. 2495–2504, 2014.

[271] A. N. Long and S. Dagogo-Jack, "Comorbidities of Diabetes and Hypertension: Mechanisms and Approach to Target Organ Protection," *Journal of Clinical Hypertension*, vol. 13, no. 4. pp. 244–251, 2011.

[272] J. Hippisley-Cox, C. Coupland, J. Robson, A. Sheikh, and P. Brindle, "Predicting risk of type 2 diabetes in England and Wales: prospective derivation and validation of QDScore," *BMJ*, vol. 338, no. 2, 2009.

[273] S. J. Griffin, P. S. Little, C. N. Hales, A. L. Kinmonth, and N. J. Wareham, "Diabetes risk score: towards earlier detection of type 2 diabetes in general practice," *Diabetes Metab Res Rev*, vol. 16, no. 3, pp. 164–171, 2000.

[274] J. Lindström and J. Tuomilehto, "The diabetes risk score: A practical tool to predict type 2 diabetes risk," *Diabetes Care*, vol. 26, no. 3, pp. 725–731, 2003.

[275] L. J. Gray, N. A. Taub, K. Khunti, E. Gardiner, S. Hiles, D. R. Webb, B. T. Srinivasan, and M. J. Davies, "The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting," *Diabet. Med.*, vol. 27, no. 8, pp. 887–895, 2010.

[276] D. Noble, R. Mathur, T. Dent, C. Meads, and T. Greenhalgh, "Risk models and scores for type 2 diabetes: systematic review," *BMJ*, vol. 343, no. d7163, 2011.

[277] N. Waugh, D. Shyangdan, S. Taylor-Phillips, G. Suri, and B. Hall, "Screening for type 2 diabetes: A short report for the National Screening Committee," *Health Technol. Assess. (Rockv).*, vol. 17, no. 35, pp. 1–89, 2013.

[278] B. J. Gray, R. M. Bracken, D. Turner, K. Morgan, M. Thomas, S. P. Williams, M. Williams, S. Rice, and J. W. Stephens, "Different type 2 diabetes risk assessments predict dissimilar numbers at 'high risk': A retrospective analysis of diabetes risk-assessment tools," *Br. J. Gen. Pract.*, vol. 65, no. 641, pp. 852–860, 2015.

[279] J. Howick, J. W. L. Cals, C. Jones, C. P. Price, A. Plüddemann, C. Heneghan, M. Y. Berger, F. Buntinx, J. Hickner, W. Pace, T. Badrick, A. Van den Bruel, C. Laurence, H. C. van Weert, E. van Severen, A. Parrella, and M. Thompson, "Current and future use of point-of-care tests in primary care: an international survey in Australia, Belgium, The Netherlands, the UK and the USA.," *BMJ Open*, vol. 4, no. 8, p. e005611, 2014.

[280] A. St John and C. P. Price, "Existing and Emerging Technologies for Point-of-Care Testing.," *Clin. Biochem. Rev.*, vol. 35, no. 3, pp. 155–167, 2014.

[281] P. B. Luppa, C. Müller, A. Schlichtiger, and H. Schlebusch, "Point-of-care testing (POCT): Current techniques and future perspectives," *TrAC - Trends in Analytical Chemistry*, vol. 30, no. 6. pp. 887–898, 2011.

[282] J. Y. Kim and K. Lewandrowski, "Point-of-Care Testing Informatics," *Clinics in Laboratory Medicine*, vol. 29, no. 3. pp. 449–461, 2009.

[283] M. Ahmad, A. Kamboh, and A. Khan, "Non-invasive blood glucose monitoring using near-infrared spectroscopy," *EDN Network*, 2013. [Online]. Available: http://www.edn.com/design/medical/4422840/Non-invasive-blood-glucose-monitoring-using-near-infrared-spectroscopy. [Accessed: 10-Dec-2015].

[284] T. D. Pigott, "A Review of Methods for Missing Data," *Educ. Res. Eval.*, vol. 7, no. 4, pp. 353–383, 2001.

[285] D. Westreich, "Berkson's bias, selection bias, and missing data," *Epidemiology*, vol. 23, no. 1, pp. 159–164, 2012.

[286] T. A. Holt, D. Stables, J. Hippisley-Cox, S. O'Hanlon, and A. Majeed, "Identifying undiagnosed diabetes: Cross-sectional survey of 3.6 million patients' electronic records," *Br. J. Gen. Pract.*, vol. 58, no. 548, pp. 192–196, 2008.

[287] T. A. Holt, C. L. Gunnarsson, P. A. Cload, and S. D. Ross, "Identification of undiagnosed diabetes and quality of diabetes care in the United States: cross-sectional study of 11.5 million primary care electronic records.," *C. open*, vol. 2, no. 4, pp. E248–E255, 2014.

[288] Y. Zhang, T. a. Holt, and N. Khovanova, "A data driven nonlinear stochastic model for blood glucose dynamics," *Comput. Methods Programs Biomed.*, vol. 125, pp. 18–25, 2015.

[289] B. N. Khandalavala, A. Rojanala, J. A. Geske, J. B. Koran-Scholl, and T. P. Guck, "Obesity bias in primary care providers," *Fam. Med.*, vol. 46, no. 7, pp. 532–535, 2014.

[290] P. Royston and I. White, "Multiple imputation by chained equations (MICE): Implementation in Stata," *J. Stat. Softw.*, vol. 45, no. 4, pp. 1–20, 2009.

[291] R. W. Brause, "Medical Analysis and Diagnosis by Neural Networks," *Proceeding ISMDA '01 Proc. Second Int. Symp. Med. Data Anal.*, pp. 1–13, 2001.

[292] M. LeBlanc and J. Crowley, "Relative risk trees for censored survival data," *Biometrics*, vol. 48, no. 2, pp. 411–415, 1992.

[293] M. S. Rahman, G. Ambler, B. Choodari-Oskooei, and R. Z. Omar, "Review and evaluation of performance measures for survival prediction models in external validation settings," *BMC Med. Res. Methodol.*, vol. 17, no. 1, p. 60, 2017.

[294] D. W. Scott, "On optimal and data-based histograms," *Biometrika*, vol. 66, no. 3, pp. 605–610, 1979.

[295] M. Dorofki, A. H. Elshafie, O. Jaafar, and O. A. Karim, "Comparison of Artificial Neural Network Transfer Functions Abilities to Simulate Extreme Runoff Data," *Int. Conf. Environ. Energy Biotechnol.*, vol. 33, pp. 39–44, 2012.

[296] W. Duch and N. Jankowski, "Transfer functions: hidden possibilities for better neural networks," *9th Eur. Symp. Artif. Neural Networks*, pp. 81–94, 2001.

[297] L. M. Rios and N. V. Sahinidis, "Derivative-free optimization: A review of algorithms and comparison of software implementations," in *Journal of Global Optimization*, 2013, vol. 56, no. 3, pp. 1247–1293.

[298] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[299] J. a. Nelder, R. Mead, B. J. a Nelder, and R. Mead, "A simplex method for function minimization," *Comput. J.*, vol. 7, pp. 308–313, 1964.

[300] W.-C. Yeh, "New parameter-free simplified swarm optimization for artificial neural network training and its application in the prediction of time series.," *IEEE Trans. neural networks Learn. Syst.*, vol. 24, no. 4, pp. 661–665, 2013.

[301] M. a Carreira-Perpiñán and G. E. Hinton, "On Contrastive Divergence Learning," *Artif. Intell. Stat.*, p. 17, 2005.

[302] R. S. Sexton, R. E. Dorsey, and J. D. Johnson, "Toward global optimization of neural networks: A comparison of the genetic algorithm and backpropagation," *Decis. Support Syst.*, vol. 22, no. 2, pp. 171–185, 1998.

[303] P. J. Werbos, "Neural networks and the experience and cultivation of mind," *Neural Networks*, vol. 32, pp. 86–95, 2012.

[304] E. F. Moreira, Miguel, "Neural Networks with Adaptive Learning Rate and Momentum Terms," *Tech. Rep. 95*, vol. 4, pp. 1–29, 1995.

[305] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Networks*, vol. 12, no. 1. pp. 145–151, 1999.

[306] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artif. Intell.*, vol. 136, no. 2, pp. 215–250, 2002.

[307] Å. Björck, *Numerical methods for least squares problems*. Philadelphia: SIAM, 1996.

[308] R. Rojas, "Neural networks: a systematic introduction," *Neural Networks*, Berlin: Springer-Verlag, p. 502, 1996.

[309] P. Borkar, M. V. Sarode, and L. G. Malik, "Employing Speeded Scaled Conjugate Gradient Algorithm for Multiple Contiguous Feature Vector Frames: An Approach for Traffic Density State Estimation," *Procedia Comput. Sci.*, vol. 78, pp. 740–747, 2016.

[310] G. Yuan, "Modified nonlinear conjugate gradient methods with sufficient descent property for large-scale optimization problems," *Optim. Lett.*, vol. 3, no. 1, pp. 11–21, 2009.

[311] A. E. Kostopoulos and T. N. Grapsa, "Self-scaled conjugate gradient training algorithms," *Neurocomputing*, vol. 72, no. 13–15, pp. 3000–3019, 2009.

[312] R. Fletcher and C. M. Reeves, "Function minimization by conjugate gradients," *Comput. J.*, vol. 7, no. 2, pp. 149–154, 1964.

[313] N. Andrei, "A scaled nonlinear conjugate gradient algorithm for unconstrained optimization," *Optimization*, vol. 57, no. 4, pp. 549–570, 2008.

[314] S. Babaie-Kafaki, "Two modified scaled nonlinear conjugate gradient methods," *J. Comput. Appl. Math.*, vol. 261, pp. 172–182, 2014.

[315] T. O. Kvålseth, "Cautionary Note about R 2," *Am. Stat.*, vol. 39, no. 4, pp. 279–285, 1985.

[316] F. E. Harrell, K. L. Lee, and D. B. Mark, "Prognostic/Clinical Prediction Models: Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors," in *Tutorials in Biostatistics, Statistical Methods in Clinical Studies*, vol. 1, 2005, pp. 223–249.

[317] P. Royston and W. Sauerbrei, "A new measure of prognostic separation in survival data," *Stat. Med.*, vol. 23, no. 5, pp. 723–748, 2004.

[318] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", *Data Mining and Knowledge Discovery*, vol. 2, no. 2, 121-167, 1998

Appendices

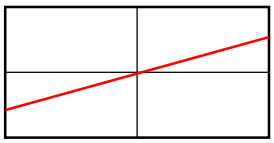
Appendix A. Neural network: extended methodology

A 1 Perceptron transfer functions

The transfer function (TF) of a neuron represents the relation between its input and output in terms of spatial or temporal frequency [87,295]. The definitions of common TFs in feedforward NNs and their first order derivatives are provided in Figure A.1. The list is by no means exhaustive: smooth rectifier (ReLU) and its modifications, logit, probit, complimentary log-log functions could be used in NNs if effective in a given application [228,295,296].

Step-like functions produce highly efficient *logical neurons*, but their discontinuous derivatives prohibit their use with gradient-based training algorithms [296]. On the other hand, sigmoidal TFs result in *graded response neurons* with differentiable output. Two sigmoidal TFs were used for the final NN models developed in this research, which are $logsig(x) = \frac{1}{1+e^{-x}}$ and $tansig(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$. The sigmoidal functions were particularly suitable for NN hidden layers as they cater for diverse behaviours: nearly linear in the vicinity of zero, nearly constant in the saturation range, and a curvilinear in the transition zone [17]. The sigmoidal TF also referred to as 'squashing' functions, due to their ability to take a real-valued input and return the output in a finite interval: [0;1] for $logsig(x)$ and [-1;1] for $tansig(x)$.

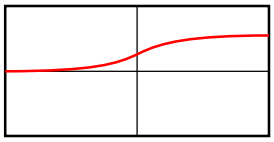
Linear



$$f(x) = kx$$

$$f'(x) = k$$

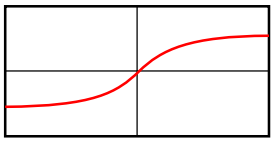
Log-sigmoid



$$f(x) = \frac{1}{1+e^{-x}}$$

$$f'(x) = \frac{1}{1+e^{-x}}\left(1 - \frac{1}{1+e^{-x}}\right)$$

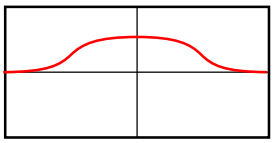
Tan-sigmoid



$$f(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

$$f'(x) = 1 - \left(\frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}\right)^{2}$$

Radial basis



$$f(x) = e^{-x^{2}}$$

$$f'(x) = -2xe^{-x^{2}}$$

Soft-max

$$f(x,i) = \frac{e^{x}{}_{i}}{\sum_{1}^{n} e^{x}{}_{i}}$$

$$f'(x,i) = \frac{e^{x}{}_{i}}{\sum_{1}^{n} e^{x}{}_{i}}\left(1 - \frac{e^{x}{}_{i}}{\sum_{1}^{n} e^{x}{}_{i}}\right)$$

Figure A.1 Common perceptron transfer functions

A 2 Backpropagation

The principles underlying backpropagation have existed since 1960s, but the method was not formalised in the context of NNs until 1986. Experiments by Rumelhart, Hinton and Williams demonstrated that backpropagation could be used to train practical multi-layer networks [100]. Stuart Dreyfus is credited for an elegant derivation of backpropagation using chain rule only [105]. The early neural networks used discrete outputs, barring the use of derivative-based methods, and it was not until LeCun [102] and Rumelhart et al. [100] overcame this problem by replacing the binary-output

neurons with those using sigmoidal (smooth) outputs. The forward and backward passes of backpropagation are described below.

Forward pass. For the following examples, input variable x_j is a scalar that refers to a single observation. This network is shown in the diagram in Figure A.2, for which the intermediate outputs are defined as follows:

$$s_k = x_j w_{jk} + b_k \qquad eq.\ \text{A.1}$$

$$p_k = f_k(s_k) \qquad eq.\ \text{A.2}$$

$$s_l = p_k w_{kl} + b_l \qquad eq.\ \text{A.3}$$

$$y = f_l(s_l) \qquad eq.\ \text{A.4}$$

where s_k and s_l represent output of the summation operator in layers k and l, respectively, p_k is the output of the f_k transfer function with input s_k and y is the output of the f_l transfer function. Combining the equations A.1 – A.4, the output y computed in the forward pass is determined as:

$$y = f_l\big(f_k\big(x_j w_{jk} + b_k\big)w_{kl} + b_l\big) \qquad eq.\ \text{A.5}$$

Backward pass. The cost function between t and y was used as a basis for finding the weighted contributions to total error by each weight and bias in the network. This was achieved by using partial derivatives at each step of the NN structure in a chain rule, as illustrated in Figure A.2.

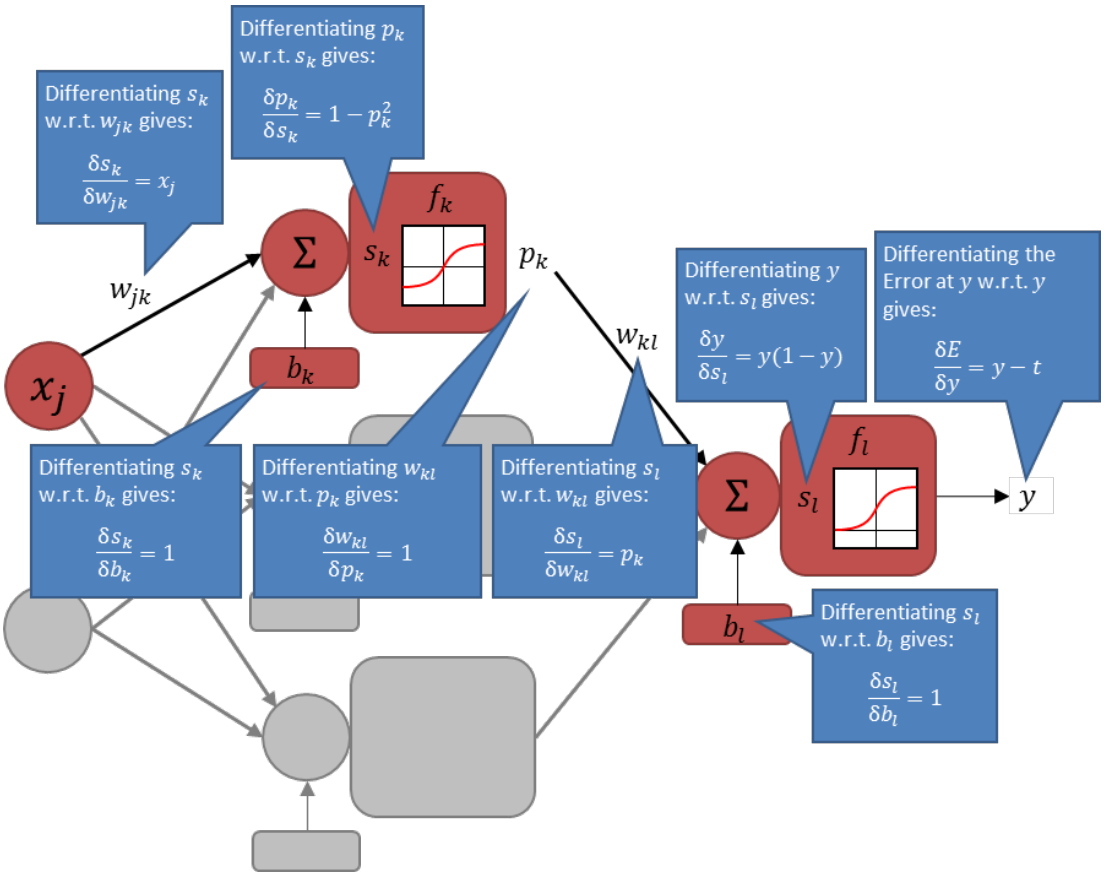


Figure A.2 Backpropagation: forward and backward passes

To determine the gradient of the cost function for w_{kl}, the value of $\frac{\delta E}{\delta w_{kl}}$ must be determined. Using the chain rule to find this derivative gives:

$$\frac{\delta E}{\delta w_{kl}} = \frac{\delta E}{\delta y} \times \frac{\delta y}{\delta s_l} \times \frac{\delta s_l}{\delta w_{kl}} \qquad eq.\ \text{A.6}$$

In order to find the gradient of the cost function w.r.t. bias b_l, the chain rule becomes:

$$\frac{\delta E}{\delta b_l} = \frac{\delta E}{\delta y} \times \frac{\delta y}{\delta s_l} \times \frac{\delta s_l}{\delta b_l} \qquad eq.\ \text{A.7}$$

Let the cost function represent a squared-error function:

$$E = \frac{1}{2}(t - y)^2$$

Let f_l represent a logarithmic sigmoid transfer function:

$$f_l = \frac{1}{1 + e^{s_l}}$$

Differentiating each term of the chain rule (equations A.6 and A.7):

$$\frac{dE}{dy} = \frac{d(\frac{1}{2}(t-y)^2)}{dy} = -\frac{2}{2}(t-y) = y - t$$

$$\frac{dy}{ds_l} = \frac{d(\frac{1}{1+e^{s_l}})}{ds_l} = \frac{1}{1+e^{s_l}} - \left(\frac{1}{1+e^{s_l}}\right)^2 = y(1-y)$$

$$\frac{ds_l}{dw_{kl}} = \frac{d(p_k w_{kl} + b_l)}{dw_{kl}} = p_k$$

$$\frac{ds_l}{db_l} = \frac{d(p_k w_{kl} + b_l)}{db_l} = 1$$

Expanding the chain rule (equations A.6 and A.7):

$$\frac{\delta E}{\delta w_{kl}} = p_k y(1-y)(y-t) \qquad eq.\ \text{A.8}$$

$$\frac{\delta E}{\delta b_l} = y(1-y)(y-t) \qquad eq.\ \text{A.9}$$

In order to extend this method for finding the gradients of the cost function w.r.t. w_{jk} and b_k, one must expand the chain rule to include the additional terms $\frac{\delta w_{kl}}{\delta p_k}, \frac{\delta p_k}{\delta s_k}, \frac{\delta s_k}{\delta w_{jk}}$ and $\frac{\delta w_{kl}}{\delta p_k}$, $\frac{\delta p_k}{\delta s_k}, \frac{\delta s_k}{\delta b_k}$ respectively:

$$\frac{\delta E}{\delta w_{jk}} = \frac{\delta E}{\delta w_{kl}} \times \frac{\delta w_{kl}}{\delta p_k} \times \frac{\delta p_k}{\delta s_k} \times \frac{\delta s_k}{\delta w_{jk}} \qquad eq.\ \text{A.10}$$

$$\frac{\delta E}{\delta b_k} = \frac{\delta E}{\delta w_{kl}} \times \frac{\delta w_{kl}}{\delta p_k} \times \frac{\delta p_k}{\delta s_k} \times \frac{\delta s_k}{\delta b_k} \qquad eq.\ \text{A.11}$$

Let f_k represent a hyperbolic tangent sigmoid transfer function:

$$f_k = \frac{e^{s_k} - e^{-s_k}}{e^{s_k} + e^{-s_k}}$$

Differentiating each additional term of the chain rule (equations A.10 and A.11):

$$\frac{\delta w_{kl}}{\delta p_k} = 1$$

$$\frac{\delta p_k}{\delta s_k} = \frac{d(\frac{e^{s_k} - e^{-s_k}}{e^{s_k} + e^{-s_k}})}{ds_k} = 1 - \left(\frac{e^{s_k} - e^{-s_k}}{e^{s_k} + e^{-s_k}}\right)^2 = 1 - {p_k}^2$$

$$\frac{\delta s_k}{\delta w_{jk}} = \frac{d(x_j w_{jk} + b_k)}{dw_{jk}} = x$$

$$\frac{\delta s_k}{\delta b_k} = \frac{d(x_j w_{jk} + b_k)}{db_k} = 1$$

Expanding the chain rule (equations A.10 and A.11):

$$\frac{\delta E}{\delta w_{jk}} = p_k x_j y(1-y)(y-t)(1-{p_k}^2) \qquad eq.\ \text{A.12}$$

$$\frac{\delta E}{\delta b_k} = p_k y(1-y)(y-t)(1-{p_k}^2) \qquad eq.\ \text{A.13}$$

These equations allowed determination of the weighted contribution of each NN parameter to the resulting error for a given sample, which are required when calculating the weight update.

Optimisation algorithms (also referred to as *solution algorithms*) are used to determine the required update to the network parameters from the cost function derivative as an

input variable. Typically, backpropagation NNs use a gradient descent method to find the optimal solution, i.e. minimise the cost function by considering the differential equation of error with respect to weight values. This involves, for instance, a simple proportional response, such as in the *gradient descent* (Section A.2.1), or a more complex, adaptive optimisation, such as in *Levenberg-Marquardt algorithm* (Section A.2.2), or *scaled conjugate gradient (SCG) method* (Section A.2.3).

Multiple alternatives to backpropagation exist for NN training [297]. These include now-obsolete techniques such as simulated annealing [298] and Nelder-Mead simplex method [299], population-based training such as evolutionary algorithms and particle swarm optimisation [300], and probabilistic methods such as radial basis functions and contrastive divergence [301]. Extensive work contrasting these stochastic and gradient-free optimisation methods with backpropagation is described in the literature [302], with the general view holding that any appropriately-tuned method can outperform others in a given application. In comparison to backpropagation, where known output values are used to direct the NN parameter adjustment, the stochastic optimisation disregards this information and attempts random changes to the parameters, often rendering them impractical for real-world applications. For those reasons, backpropagation remains dominant in feedforward NN training [103,303].

A.2.1 Gradient descent

Gradient descent is a first-order iterative optimisation algorithm which uses the gradient of the cost function with respect to each weight in order to determine the direction and magnitude of the required change to the present weight value. The weight updates can then be calculated as a step in the direction of decreasing error ("descending" the error gradient), usually with a magnitude which is a fraction of the current gradient.

The learning rule specifies the weight updates with regards to the gradient derived by gradient descent. For a gradient descent approach, the learning rule at a specific weight w_{jk} is given by the product of the neuron error, the first derivative of the activation function and the input to the weight:

$$\Delta w_{jk} = -\alpha \frac{\delta E_k}{\delta w_{kl}} p_k s_k x_j$$

where Δw_{jk} is the weight update at weight w_{jk}, α is the learning rate, δE_k is the gradient of the error function with respect to w_{jk}, s_k is the sum of weighted inputs at neuron k, and x_j is the input to the weight w_{jk}.

One limitation of the gradient descent algorithm is the possibility of training the weights and biases to reach a local minimum, hence constraining the potential gains of the learning process. This occurs as a result of the nature of the gradient descent, as it simply seeks to minimise the size of the error, without general knowledge of the error function outside of the near vicinity in the variable space. The risk of optimising for local minima can be reduced by selecting appropriate learning rates [87,304].

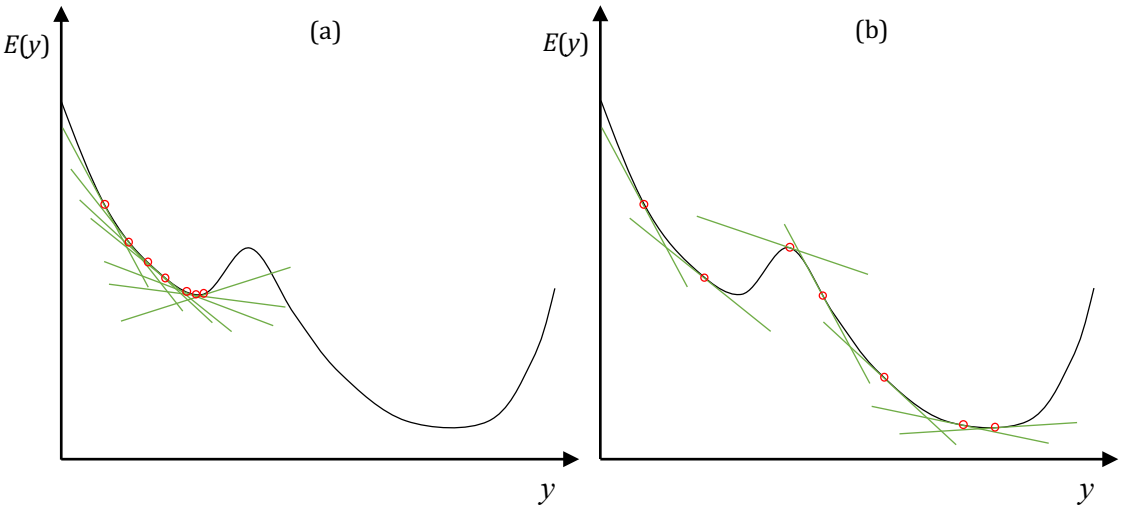


Figure A.3 Effect of changing learning rate on saddle point local minima

Figure A.3 (a) illustrates the potential problem of reaching a local minimum when using an inadequate learning rate, while Figure A.3 (b) contrasts the convergence towards a lower minimum when using a different learning rate. It is also worth noting that a high learning rate may result in an unstable oscillating response during training, while a low learning rate may result in a very slow learning process.

Similarly to learning rate, initial weight and bias values also influence the ability of the NN to converge to minimum error solutions, as they can cause the learning algorithm to optimise local minimums and saddle points. Therefore, the success of gradient descent backpropagation algorithm depends on arbitrary user-defined learning rate and initial weight and bias values.

One modification of gradient descent applies the use of a "momentum term" in the weight update equation, which essentially transforms an online-learning approach to a mini-batch approach using a moving average [305]. More recent work has demonstrated the merits of employing a gradient descent learning algorithm with variable learning rate [306]. This approach not only optimises the trade-off between a fast learning process and an effective optimisation, but also reduces the dependency on selecting adequate starting weight values [304].

A.2.2 Gauss-Newton and Levenberg-Marquardt algorithms

Levenberg-Marquardt algorithm operates by finding a solution for a non-linear least squares problem. The algorithm alternates between gradient descent and Gauss-Newton methods depending on the outcome of previous iterations using a variable learning rate [112]. Gauss-Newton method can be described by considering a nonlinear least-squares optimisation problem:

$$E = \frac{1}{2}\sum_{i=1}^{n} {r_i}^2(\beta)$$

where n is to the number of observations, $r_i(\beta)$ is the error function between the target t_i and the predicted variable, and β is the vector of model parameters $\beta = [\beta_1, \beta_2, \dots, \beta_m]$, where the model comprises m parameters and E represents the summation of squared errors. The error can be expressed as follows:

$$r_i(\beta) = t_i - f(x_i, \beta) \qquad eq.\ \text{A.14}$$

where $f(x_i, \beta)$ represents the predicted variable resulting from the model. To identify the set of model parameters that will minimise the squared error, an iterative approach can be adopted, while taking an initial assumption for the parameters $\beta^{(0)}$, and modifying this by a pre-determined step. The step is determined using a Newton method of finding minima to a function, that uses a Taylor series quadratic approximation [307].

The parameter update equation takes the form:

$$\beta^{(s+1)} = \beta^{(s)} - gH^{-1} \qquad eq.\ \text{A.15}$$

where H is the Hessian matrix of the error, composed of the second-order partial derivatives of the error function, and g is the first-order derivative vector. Shown below is the Hessian matrix specific to the error function defined in $eq.$ A.14:

$$H = \begin{bmatrix} \sum_{i=1}^{n}\left(\frac{\partial r_i}{\partial \beta_1}\right)^2 + \frac{\partial^2 r_i}{\partial \beta_1^2} & \sum_{i=1}^{n}\frac{\partial r_i}{\partial \beta_1}\frac{\partial r_i}{\partial \beta_2} + \frac{\partial^2 r_i}{\partial \beta_1 \partial \beta_2} & \cdots & \sum_{i=1}^{n}\frac{\partial r_i}{\partial \beta_1}\frac{\partial r_i}{\partial \beta_m} + \frac{\partial^2 r_i}{\partial \beta_1 \partial \beta_m} \\ \sum_{i=1}^{n}\frac{\partial r_i}{\partial \beta_2}\frac{\partial r_i}{\partial \beta_1} + \frac{\partial^2 r_i}{\partial \beta_2 \partial \beta_1} & \sum_{i=1}^{n}\left(\frac{\partial r_i}{\partial \beta_2}\right)^2 + \frac{\partial^2 r_i}{\partial \beta_2^2} & & \sum_{i=1}^{n}\frac{\partial r_i}{\partial \beta_2}\frac{\partial r_i}{\partial \beta_m} + \frac{\partial^2 r_i}{\partial \beta_2 \partial \beta_m} \\ \vdots & & \ddots & \vdots \\ \sum_{i=1}^{n}\frac{\partial r_i}{\partial \beta_m}\frac{\partial r_i}{\partial \beta_1} + \frac{\partial^2 r_i}{\partial \beta_m \partial \beta_1} & \sum_{i=1}^{n}\frac{\partial r_i}{\partial \beta_m}\frac{\partial r_i}{\partial \beta_2} + \frac{\partial^2 r_i}{\partial \beta_m \partial \beta_2} & \cdots & \sum_{i=1}^{n}\left(\frac{\partial r_i}{\partial \beta_m}\right)^2 + \frac{\partial^2 r_i}{\partial \beta_m^2} \end{bmatrix} \qquad eq.\ \text{A.16}$$

$$g = \left(\sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \beta_1}, \sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \beta_2}, \ldots, \sum_{i=1}^{n} r_i \frac{\partial r_i}{\partial \beta_m} \right)$$

For small steps, the second-order terms in the Hessian matrix can be ignored, and both the Hessian matrix and the gradient vector can be expressed in terms of the respective Jacobians:

$$H = J^T J$$

$$g = J^T r$$

where the Jacobian matrix has the format:

$$J = \begin{bmatrix} \frac{\partial r_1}{\partial \beta_1} & \frac{\partial r_1}{\partial \beta_2} & \cdots & \frac{\partial r_1}{\partial \beta_m} \\ \frac{\partial r_2}{\partial \beta_1} & \frac{\partial r_2}{\partial \beta_2} & & \frac{\partial r_2}{\partial \beta_m} \\ \vdots & & \ddots & \vdots \\ \frac{\partial r_n}{\partial \beta_1} & \frac{\partial r_n}{\partial \beta_2} & \cdots & \frac{\partial r_n}{\partial \beta_m} \end{bmatrix}$$

We can therefore rewrite the step change equation (A.14):

$$\beta^{(s+1)} = \beta^{(s)} - (J^T J)^{-1} J^T r \qquad eq. \text{A.17}$$

Levenberg Marquardt is effectively a modification of the Gauss-Newton algorithm, and uses the parameter λ to blend gradient descent and Gauss-Newton methods:

$$\beta^{(s+1)} = \beta^{(s)} - (J^T J + \lambda I)^{-1} J^T r \qquad eq. \text{A.18}$$

If the error is reduced following a weight update, λ is decreased to increase the contribution of Gauss-Newton, therefore increasing the sensitivity of the update algorithm. If the error increases after a weight update, λ is increased and the weight

update reversed, allowing gradient descent to more rapidly guide the weight update. In this manner, Levenberg Marquardt is able to combine the sensitivity of a Gauss-Newton method with the speed of convergence of gradient descent, greatly improving training efficiency, as demonstrated by Rojas [308].

A.2.3 Conjugate gradient method

Conjugate gradient methods are used to calculate the global minimum for a function of the form $Ax = b$, where A is a symmetric positive definite n-by-n matrix, and x and b are vectors of size n [115,309]. Similar to gradient descent, the conjugate gradient method takes an iterative approach to finding the minimum of a function, but additionally uses the concept of *conjugate vectors* to find an optimal step size [310,311]. A pair of non-zero vectors u and v, is said to be conjugate when:

$$u^T A v = 0$$

The first iteration in the conjugate gradient method is performed in a manner similar to gradient descent, with direction proportional to the steepest gradient at the first estimated solution, and magnitude optimised by a line search. The direction p of all subsequent iterations, however, must be orthogonal to all previous steps, and for step k is given by the formula:

$$p_k = -g_k + \beta_k p_{k-1} \qquad eq.\ \text{A.19}$$

where g_k is the gradient at the current estimate, β_k is an adaptive constant that ensures the orthogonality of the gradient [115]. The equation to calculate β_k varies across different conjugate gradient algorithms; the one used in this study is Fletcher-Reeves approach [312], in which β_k is given by:

$$\beta_k = \frac{{g_k}^T g_k}{{g_{k-1}}^T g_{k-1}} \qquad eq.\ \text{A.20}$$

The conjugate gradient method relies on the assumption that the error in the neighbourhood of the current estimate $E(w_k)p_k$ can be approximated by a quadratic equation. The step size is then calculated using a line search algorithm, which is computationally intensive. A modification, called scaled conjugate gradient (SCG), was introduced by Møller [115], with the aim of combining the conjugate gradient with a model trust region, similar to Levenberg-Marquardt. This approach uses the following approximation to the second order terms in the Hessian matrix of the error function $E(w_k)p_k$ with respect to weights w_k :

$$\partial^2 E(w_k)p_k \approx \frac{\partial E(w_k + \sigma_k p_k) - \partial E(w_k)}{\sigma_k} + \lambda_k p_k \qquad eq.\ \text{A.21}$$

where σ_k and λ_k are arbitrary small positive constants. This approximation reduces the complexity of the computation, increasing the training efficiency [313,314]. The proof and detail of this method are presented by Møller [115].

Appendix B. Performance criteria

The regression factor R compares the sum of squares due to error and the total sum of squares, and, in its standard form, is given by:

$$R = \sqrt{1 - \frac{\sum_{i=1}^{n}(t_i - y_i)^2}{\sum_{i=1}^{n}(t_i - \bar{y})^2}}$$

where t_i is the target output, y_i is the output predicted by the model, $\bar{y}$ is the mean of y_i and n is a total number of samples [315]. R takes values between 0 and 1, where R = 1 corresponds to a perfect fit, signifying that the entire variance in the dependent output variable can be explained by the regression equation. R greater than 0.6 defines statistically significant performance, i.e. $R_{all} \geq 0.6, R_{tr} \geq 0.6, R_{val} \geq 0.6$, and $R_{test} \geq 0.6$ [84].

$RMSE$, provides the same information as R, but is expressed in terms of the *absolute* difference between model predictions and targets, making it particularly useful for visualising the error in the units of the output variable:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(t_i - y_i)^2}$$

In classification problems, where t_i is dichotomous (0 or 1), the model accuracy is expressed by the proportion of observations with correctly predicted class [182]. With NN and DT classifiers, the predicted outcome y_i is a continuous real-valued number (between 0 and 1) that describes a probability of each class. In order to dichotomise y_i into a binary class label, a cut-off is applied at a given discrimination threshold (0.5 default value). *Receiver operating characteristic* (ROC) curve [182] depicts Sp versus $1 - Sn$ at various thresholds, where each point represents a different trade-off between FP

and FN predictions. This trade-off can be quantified as *cost ratio*, which is defined by the gradient of the line tangent to the ROC curve. The *area under the ROC curve* (AUC) represents the performance averaged across all cost ratios. On the unit ROC space, a perfect prediction would yield an AUC of 1.0. A random coin flipping would result in points along the diagonal and the corresponding AUC of 0.5. Tracing back its origins to World War II, where it was developed to model FP and FN radar detections, ROC is now a standard measure of discrimination in medicine and clinical science, and its use is becoming increasingly popular in the ML community [182].

In survival modelling the true outcome t_i may not be known for some patients due to censoring. A generalization of the AUC for assessing the discriminating ability of a survival model is known as *concordance* – a rank correlation between the *predicted risk* y_i and the *observed survival times* T for all comparable pairs [293]. For time-to-event outcomes, pairs of patients (i, j) are *comparable* (useable) if $T_i \not\succ T_j$, at least one of the patients is uncensored. The pair is said to be *concordant* when the patient with the lowest predicted risk y out survives the other, i.e. has the longer survival time T. The most popular concordance measure in biostatistics and prognostic modelling is known as Harrell's $C - index$, which defines the probability of concordant pairs in all usable pairs [316] as follows:

$$C\text{-index} = P\left(y_i > y_j \middle| T_i < T_j\right)$$

Although it had been recently suggested that Uno's or Gonen & Heller's concordance measures are more suitable for applications with high proportion of censoring [293], Harrell's C-index was adopted in Chapter 6 in order to ensure consistency of comparison with the existing studies, as specified by the collaboration protocol [90].

Another measure of discrimination popular in survival analysis is *Royston and Sauerbrei's D* [317]. The D statistic quantifies the relative gain in prognostic separation between two equally-sized groups of patients with lowest and highest predicted risk scores y defined at the median value. This is achieved by ordering y, calculating the expected normal order statistics, scaling them by factor $k = \sqrt{8/\pi}$ and performing auxiliary regression on the scaled values [317]. What makes Royston and Sauerbrei's D particularly appealing for assessing Cox PH survival models, is that it provides an indication of the overall model *log hazard ratio*.

An R^2-like form could be achieved by transforming D into R_D^2 which, in turn, is interpreted as the proportion of prognostic separation explained by the model:

$$R_D^2 = \frac{D^2/k^2}{\sigma^2 + D^2/k^2}$$

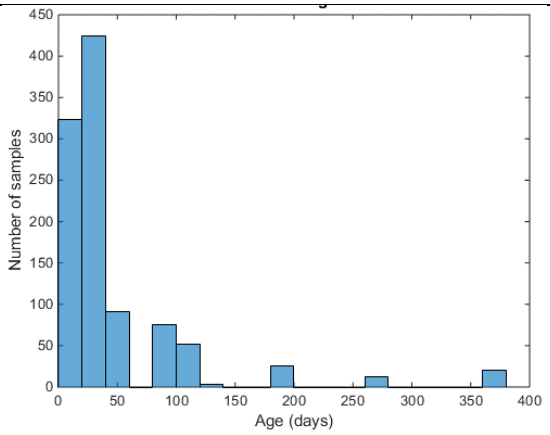
$$\text{where } \sigma = \begin{cases} 1 \text{ for lognormal models} \\ \pi^2/3 \text{ for loglogistic or proportional odds models} \\ \pi^2/6 \text{ for proportional hazard models} \end{cases}$$

Appendix C. Concrete compressive strength dataset

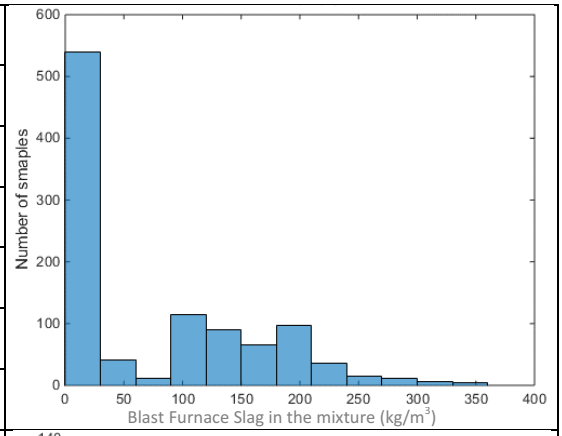
The subsets of concrete CS data were accessed through a public repository [164,165] and partitioned as follows for the 3 NN models considered in Chapter 3: (a) large-dataset NN (730 model samples and 300 tests), (b) intermediate 100-sample NN (70 model samples and 30 tests), (c) small-dataset NN (41 model samples and 15 tests). The model samples refer to the validation and training samples, combined.

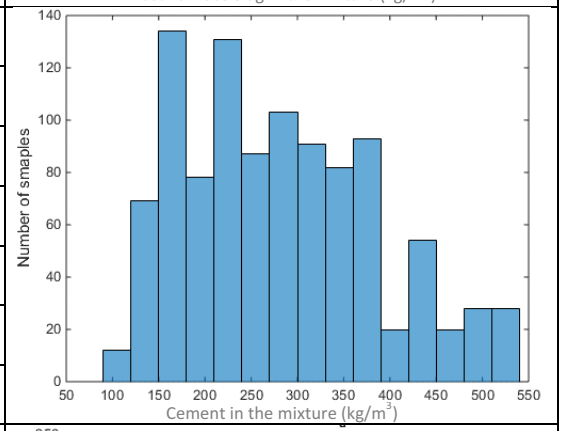
Table C.1 provides the key statistics on each model (a), (b) and (c) with the breakdown according to the model and test subsets (number of samples given in brackets). The frequency distribution histograms for each of the 9 variables x in the complete CS dataset are included in the last column.

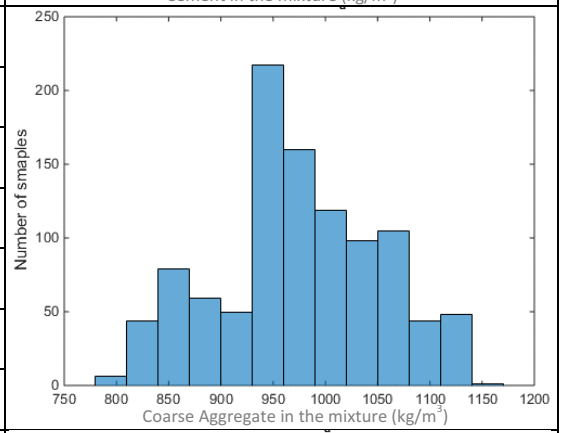
Table C.1 Concrete CS dataset statistics by individual variable

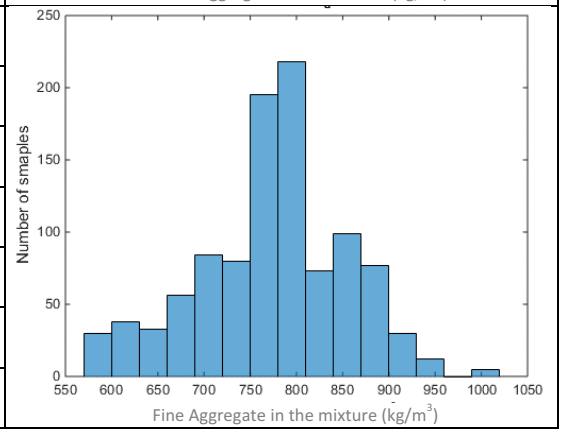
x	Model	Subset	*min*	*max*	μ	σ	Frequency distribution of the variable across all (1030) samples
Age (days)		all (1030)	1.0	365.0	45.7	63.2	
	a	model (730)	1.0	356.0	44.6	62.5	
		test (300)	3.0	365.0	48.3	64.9	
	b	model (70)	3.0	365.0	50.5	69.3	
		test (30)	3.0	365.0	41.1	64.8	
	c	model (41)	3.0	365.0	45.5	70.2	
		test (15)	3.0	180.0	43.6	46.1	

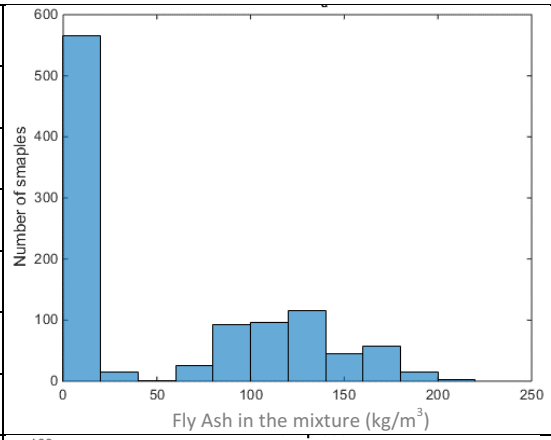
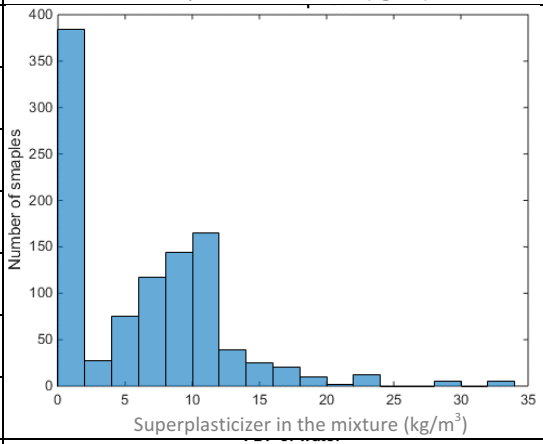
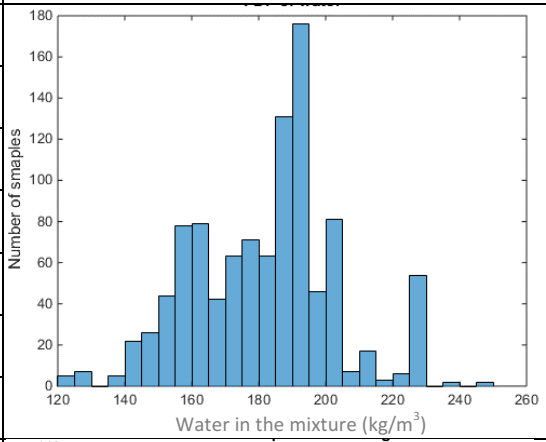
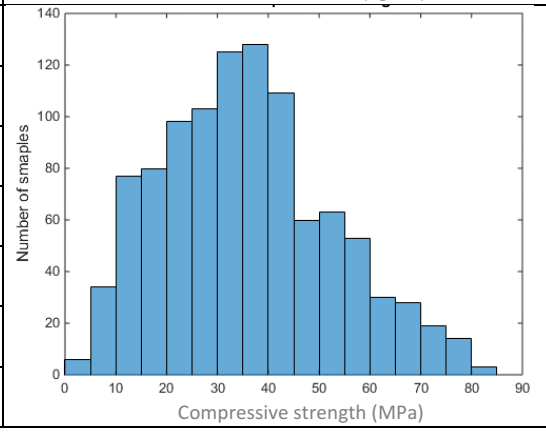
Blast Furnace Slag (kg/m³)		all (1030)	0.0	359.4	73.9	86.3
	a	model (730)	0.0	359.4	73.3	86.3
		test (300)	0.0	342.1	75.4	86.4
	b	model (70)	0.0	316.1	66.4	81.7
		test (30)	0.0	239.0	68.2	87.3
	c	model (41)	0.0	359.4	86.1	84.9
		test (15)	0.0	250.2	90.6	90.4
Cement (kg/m³)		all (1030)	102.0	540.0	281.2	104.5
	a	model (730)	102.0	540.0	279.9	103.9
		test (300)	108.3	540.0	284.2	106.1
	b	model (70)	102.0	540.0	286.8	117.1
		test (30)	145.9	525.0	290.6	115.9
	c	model (41)	140.0	500.0	285.0	111.7
		test (15)	102.0	522.0	268.4	121.1
Coarse Aggregate (kg/m³)		all (1030)	801.0	1145.0	972.9	77.8
	a	model (730)	801.0	1145.0	973.1	76.9
		test (300)	801.0	1134.0	972.6	79.8
	b	model (70)	822.0	1125.0	965.0	83.4
		test (30)	827.0	1125.0	986.4	73.6
	c	model (41)	822.0	1125.0	972.4	81.4
		test (15)	814.0	1069.0	930.8	63.3
Fine Aggregate (kg/m³)		all (1030)	594.0	992.6	773.6	80.2
	a	model (730)	594.0	992.6	774.7	79.2
		test (300)	594.0	992.6	770.8	82.6
	b	model (70)	613.0	943.1	779.1	77.9
		test (30)	594.0	896.0	767.0	74.6
	c	model (41)	594.0	943.1	772.2	76.8
		test (15)	594.0	942.0	787.6	95.1









Fly Ash (kg/m³)		all (1030)	0.0	200.1	54.2	64.0	
	a	model (730)	0.0	200.1	55.5	64.3	
		test (300)	0.0	200.1	51.0	63.2	
	b	model (70)	0.0	194.9	48.1	59.5	
		test (30)	0.0	174.2	56.4	63.6	
	c	model (41)	0.0	174.9	49.4	63.1	
		test (15)	0.0	195.0	52.9	72.4	
Superplasticizer (kg/m³)		all (1030)	0.0	32.2	6.2	6.0	
	a	model (730)	0.0	32.2	6.3	5.9	
		test (300)	0.0	32.2	5.9	6.1	
	b	model (70)	0.0	32.2	5.7	6.1	
		test (30)	0.0	32.2	7.9	7.2	
	c	model (41)	0.0	32.2	6.1	6.9	
		test (15)	0.0	18.0	6.3	6.0	
Water (kg/m³)		all (1030)	121.8	247.0	181.6	21.4	
	a	model (730)	121.8	247.0	181.2	21.0	
		test (300)	126.6	228.0	182.5	22.1	
	b	model (70)	137.8	246.9	184.7	19.4	
		test (30)	137.8	228.0	178.6	23.8	
	c	model (41)	121.8	228.0	181.0	22.2	
		test (15)	146.0	228.0	184.8	27.2	
Compressive strength (MPa)		all (1030)	2.3	82.6	35.8	16.7	
	a	model (730)	3.3	82.6	35.6	16.2	
		test (300)	2.3	81.8	36.3	17.9	
	b	model (70)	4.6	72.1	36.1	17.5	
		test (30)	12.5	75.0	36.9	15.1	
	c	model (41)	4.8	80.2	34.8	18.9	
		test (15)	7.7	75.0	35.9	19.0	

Appendix D. Bone dataset: real and surrogate data

Bone data in Table D.2 were extracted from the original study by Perilli et al. [166] using a Plot Digitiser tool [160]. Surrogate data in Table D.3 were synthesised as a random normal distribution with the mean and standard deviation of the real bone data within the same range (Section 3.4.5).

Table D.2 Trabecular bone data

Sample	SMI	Tb.Th (μm)	BV/TV (%)	Age (years)	Gender (F=1)	CS (MPa)
1	0.06	243	32.5	41.8	1	20.9
2	1.42	224	21.5	52.0	1	6.91
3	0.48	239	26.6	57.0	1	18.2
4	-0.82	212	43.5	63.9	1	9.46
5	1.22	419	17.9	64.0	1	23.1
6	0.64	223	27.6	67.1	1	19.4
7	2.10	197	9.82	68.1	1	2.76
8	0.38	367	26.9	71.5	1	18.9
9	0.80	218	15.4	74.9	1	6.49
10	0.54	314	25.0	76.0	1	17.8
11	0.30	326	32.4	87.0	1	24.2
12	-0.17	287	30.4	41.7	0	21.5
13	-0.31	284	37.0	47.9	0	16.4
14	0.04	265	38.7	49.8	0	11.1
15	0.82	241	22.7	49.8	0	26.5
16	-0.23	303	37.6	65.8	0	28.8
17	1.77	219	25.3	68.0	0	4.91
18	1.33	261	17.4	72.9	0	9.81
19	0.04	307	29.7	73.9	0	23.7
20	0.36	271	31.6	81.8	0	24.4
21	0.31	252	33.8	60.9	1	20.5
22	0.70	283	22.5	62.9	1	12.2
23	1.59	247	13.7	72.6	1	1.93
24	0.45	257	27.4	45.7	0	19.6
25	0.44	266	27.5	62.9	0	18.5
26	0.15	270	32.1	77.8	0	22.2
27	1.08	193	19.4	87.0	0	9.12
28	1.93	154	9.68	49.0	1	8.22
29	0.92	263	25.3	66.0	1	15.4
30	-0.43	299	39.7	69.9	1	23.2
31	1.04	239	21.0	73.9	1	8.15
32	-0.05	288	35.6	46.8	0	24.3
33	0.39	246	26.6	64.9	0	19.3
34	0.71	178	12.2	68.0	0	14.0
35	0.70	234	21.8	84.9	0	13.3

Table D.3 Surrogate data

Sample	SMI	Tb.Th (μm)	BV/TV (%)	Age (years)	Gender (F=1)	CS (MPa)
1	0.06	243	32.5	41.8	1	20.9
2	1.42	224	21.5	52.0	1	6.91
3	0.48	239	26.6	57.0	1	18.2
4	-0.82	212	43.5	63.9	1	9.46
5	1.22	419	17.9	64.0	1	23.1
6	0.64	223	27.6	67.1	1	19.4
7	2.10	197	9.82	68.1	1	2.76
8	0.38	367	26.9	71.5	1	18.9
9	0.80	218	15.4	74.9	1	6.49
10	0.54	314	25.0	76.0	1	17.8
11	0.30	326	32.4	87.0	1	24.2
12	-0.17	287	30.4	41.7	0	21.5
13	-0.31	284	37.0	47.9	0	16.4
14	0.04	265	38.7	49.8	0	11.1
15	0.82	241	22.7	49.8	0	26.5
16	-0.23	303	37.6	65.8	0	28.8
17	1.77	219	25.3	68.0	0	4.91
18	1.33	261	17.4	72.9	0	9.81
19	0.04	307	29.7	73.9	0	23.7
20	0.36	271	31.6	81.8	0	24.4
21	0.31	252	33.8	60.9	1	20.5
22	0.70	283	22.5	62.9	1	12.2
23	1.59	247	13.7	72.6	1	1.93
24	0.45	257	27.4	45.7	0	19.6
25	0.44	266	27.5	62.9	0	18.5
26	0.15	270	32.1	77.8	0	22.2
27	1.08	193	19.4	87.0	0	9.12
28	1.93	154	9.68	49.0	1	8.22
29	0.92	263	25.3	66.0	1	15.4
30	-0.43	299	39.7	69.9	1	23.2
31	1.04	239	21.0	73.9	1	8.15
32	-0.05	288	35.6	46.8	0	24.3
33	0.39	246	26.6	64.9	0	19.3
34	0.71	178	12.2	68.0	0	14.0
35	0.70	234	21.8	84.9	0	13.3

Appendix E. Multiple imputation in diabetes data

Imputation accuracy of BG and BMI values using MICE (Section 3.1.3) was investigated on a cohort of 14922 patients for whom the BG and BMI values were known. These samples were randomly injected with missing values in BG and BMI columns to reproduce the proportion of missing values in the original cohort of 79959 patients (Figure E.1).

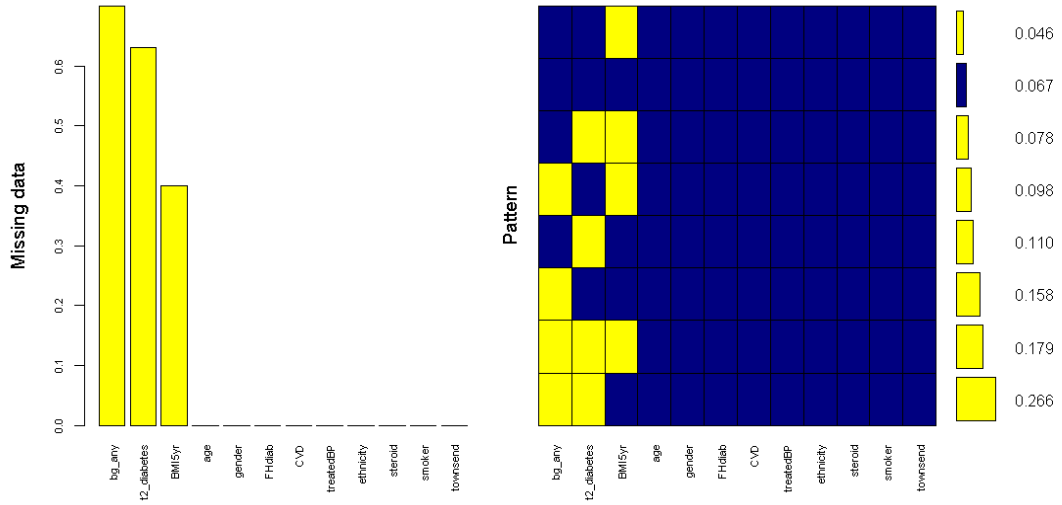


Figure E.4 Missing patterns reproduced to match the original missing proportions: 70% of BG values and 40% of BMI values are missing.

Subsequently, the deliberately introduced missing values were imputed using MICE algorithm with *m* iterations. The model included gender, age, family history of DM, CVD, hypertension, ethnicity, steroid use, smoking status, Townsend index and type 2 DM outcome.

Table E.1 indicates the $RMSE$ values for BG and BMI averaged across all imputed samples for various *m* from 1 to 100. For this particular study, the number of MICE iterations had a marginal effect on the overall imputation error, which was equal to 0.96 mmol/L for BG and 4.7 for BMI, corresponding to 19% and 18% relative error, respectively.

Table E.4 Imputation accuracy of MICE at 70% missing BG and 40% missing BMI

	$RMSE$ between actual and imputed values			
BG (mmol/L)	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	**min**	**max**
m=1	0.949	NA	0.949	0.949
m=5	0.962	0.012	0.948	0.978
m=10	0.957	0.009	0.938	0.968
m=50	0.959	0.009	0.936	0.980
m=100	0.957	0.010	0.933	0.983
BMI (kg/m²)	$\boldsymbol{\mu}$	$\boldsymbol{\sigma}$	**min**	**max**
m=1	4.703	NA	4.703	4.703
m=5	4.759	0.000	4.759	4.759
m=10	4.744	0.000	4.744	4.744
m=50	4.756	0.000	4.756	4.756
m=100	4.757	0.000	4.757	4.757

It is important to note that $RMSE$ is not the only indication of the effectiveness of imputation. As discussed in Section 3.1.3, MICE does not optimise for individual sample accuracy, but instead attempts to reproduce the overall resemblance to a complete dataset by generating multiple datasets.